Computational analysis of a candidate region for psychosis

Stéphane Ballereau

PhD

The University of Edinburgh

2004

DECLARATION

I declare that:

(a) This thesis has been composed by myself
(b) That the work is my own, except where otherwise stated

Stéphane J Ballereau

Septembre 2004

ACKNOWLEGMENTS

I would like to thank my supervisors Colin Semple, Kathryn Evans and David Porteous for their help, guidance and encouragement. I am indebted to them, and to David in particular, for giving me the opportunity to join the group and participate to the fascinating '4p' project. Together with Stéphanie Le Hellard and Stewart Morris, they made these years of training in research in a foreign country an unforgettable experience. Kathy and Stéphanie helped me understand issues in human molecular genetics and in particular positional cloning. I am grateful to Stewart for assisting me in my learning of programming, in solving technical problems and for answering daily questions. Colin played a key role in my discovering of bioinformatics and had a particularly significant impact on my training and research. I am also deeply thankful to Colin for continuing his supervision after leaving the group to start a new position.

Most of the work presented in this thesis relied on data made publicly available by the genome biology community and on the computational facilities provided by the Human Genome Mapping Project – Resource Centre founded by Medical Research Council.

Last but not least, I am thankful to my family and friends for their continuing support.

ABSTRACT

Bipolar affective disorder (BPAD) and schizophrenia are severe psychiatric diseases with a strong genetic basis. Analysis of putative susceptibility genes requires the identification of polymorphisms, such as microsatellites and Single Nucleotide Polymorphisms (SNPs), in coding and non-coding sequences. The aim of this thesis was to develop and use bioinformatic tools to direct laboratory work on a candidate region for BPAD and schizophrenia on the short arm of human chromosome 4. Specifically, the research aimed to facilitate the management and analysis of data generated by allelic association studies, and identify putative functional sequences as targets for SNP selection.

This thesis first reports the creation of an extension of the ACeDB database used by the laboratory to manage the sequence of the candidate region, and a front-end to manage and analyse raw data generated by allelic association studies based on individual genotypes and pooled DNA. It then describes various strategies to define putative non-coding functional sequences benchmarked against a set of genes with experimentally verified regulatory regions. Initially, methods for detecting known transcription factor binding sites and novel motifs were combined with tests of sequence conservation between human and mouse. This approach showed modest sensitivity and generated a high rate of false positives. Next, a purely comparative genomics approach was explored using publicly available genome sequences from human, mouse, rat, chicken, zebrafish and pufferfish. Several scoring systems for multiple alignments of genomic sequences were tested. The most successful scoring scheme detected regulatory sequences in 78% of the control genes using human, mouse and rat sequences. The inclusion of the chicken and fish sequences in the analysis reduced the number of genes that could be considered and the accuracy of all scoring systems. Predictions of the functional potential of non-coding DNA were sufficiently sensitive and specific to suggest a limited number of putative functional regions for experimental verification. This approach was therefore applied to genes in the 4p region linked to psychosis.

The work presented in this thesis fulfilled three goals towards the characterisation of candidates genes in the 4p region linked to psychosis: i) the development of a computational tool to assist ongoing laboratory work by facilitating association data management and analysis, ii) the *in silico* identification of non-coding sequences with a functional potential using fast developing comparative genomics and iii) the selection of SNPs located in these regions as candidate markers for large-scale case-control allelic association studies.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

PUBLICATIONS

- The meso-genomic era (Semple et al., 2001)
- SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis (Le Hellard et al., 2002).

Permission to include these two articles in the thesis was obtained from each publisher.

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

1.1 Identification of genes involved in complex diseases

A Mendelian trait is controlled by the genotype at a locus that is necessary and sufficient for the expression of that character. Genes causing human Mendelian traits and diseases have been discovered by positional cloning whereby a gene is identified solely on the basis that the phenotype is inherited (Morton, 1955). The chromosomal region that co-segregates with the disease in families with affected members is determined by linkage analysis (Botstein et al., 1980). A candidate gene in this region is then identified and causal variants are found in patients. Diseases for which genes have been identified using positional cloning include: Duchenne muscular dystrophy (Koenig et al., 1987), cystic fibrosis (Kerem et al., 1989), breast cancer (Miki et al., 1994; Wooster et al., 1995) and Huntington disease (Gusella et al., 1983).

Many complex diseases do not segregate in families as Mendelian traits, but are instead influenced by multiple genetic and environmental factors (Lander and Schork, 1994; Botstein and Risch, 2003). The major cause for diseases known to date is mutation in coding regions of genes which affect protein quantity and/or quality (Cargill et al., 1999). There is increasing evidence that polymorphisms in regulatory regions may increase risk of diseases (Risch and Merikangas, 1996; Collins et al., 1997; Wray et al., 2003; Knight, 2005). Some diseases have also been shown to implicate epigenetic changes.

The term epigenetics refers to the transmission of information from a cell or multicellular organism to its descendants without that information being encoded in the nucleotide sequence of genes (Grunstein, 1998; Davis and Brackmann, 2003; Sollars et al., 2003). Transmission of epigenetic factors from one generation to the next is partially stable. Epigenetic marking is mediated by DNA methylation and core histones methylation and acetylation, and creates molecular landmarks that differentiate between active and inactive chromatin. Epigenetic modifications play an important role during development and adult life via the inactivation of the X chromosome, genomic imprinting and gene silencing. At an imprinted locus, one

allele is typically transcriptionally silent but epigenetic heterogeneity is recognised between individuals (Pastinen et al., 2004). The environment may also affect epigenetic factors (Petronis and Petroniene, 2000; Wong et al., 2005). Epigenetic factors may also be involved in several complex diseases (Petronis, 2001; Wong et al., 2005) such as inflammatory bowel disease (Petronis and Petroniene, 2000) and neuropsychiatric diseases (Petronis et al., 2000; Aranyi et al., 2002; Kim et al., 2003; Petronis, 2003; Petronis et al., 2003; van Overveld et al., 2003; Abdolmaleky et al., 2004; Kan et al., 2004; Caballero and Hendrich, 2005). For example, abnormal epigenetic marking may be implicated in facioscapulohumeral muscular distrophy (FSHD; OMIM 158900) and the Rett syndrome (OMIM 158900). Most patients with FSHD carry deletions of a number of 3.3 Kb tandem repeats, termed D4Z4, located on 4q35. The repeats are linked to hypomethylation of a silencer which correlates with upregulation of nearby genes (van Overveld et al., 2003). Rett syndrome is a neurodevelopmental disorder involving impaired speech and motor skills, as well as seizures and autistic behaviour. It is caused by mutations in the gene encoding the *methyl-CpG-binding repressor protein* (*MECP2*). The consequent imperfect repression of *brain-derived neurotrophic factor* (*BDNF*) may be involved in altered synaptic plasticity causing the Rett syndrome (Caballero and Hendrich, 2005).

The genetic bases of complex traits, including human behavioural traits and psychiatric disorders, are poorly understood. Many genes implicated in complex disorders have nonetheless been identified using positional cloning (Botstein and Risch, 2003). Approaches which have revealed genes of interest include the identification of chromosomal abnormalities via cytogenetics, the study of families where disorders segregate in a Mendelian manner, linkage disequilibrium mapping (Lander and Schork, 1994; Jorde, 2000) and the use of endophenotypes, biological traits that are associated with target behavioural phenotypes, such as brain activity in response to stimuli measured by functional Magnetic Resonance Imaging (MRI) (van Rijn et al., 2005).

Cytogenetic abnormalities have proven to be valuable in the understanding of complex diseases. Several disorders in human are caused by chromosomal translocation, insertion or deletion that altered gene expression resulting in similar or identical phenotypes as those due to point mutations (Collins, 1995; McIntyre et al.,

2004; Pickard et al., 2005). Several neuro-psychiatric disorders have been reported to involve chromosomal abnormalities in or near genes, such as *dyslexia susceptibility 1 candidate 1* (*DYX1C1*) in developmental dyslexia (Taipale et al., 2003), *forkhead box P2* (*FOXP2*) in speech and language development (Lai et al., 2001; Newbury et al., 2002; Lai et al., 2003), *suppression of tumerogenicity 7* (*ST7*) and *autism susceptibility candidate 2* (*AUTS2*) in autism (Vincent et al., 2000a; Sultana et al., 2002; Vincent et al., 2002), *disrupted in schizophrenia 1* and *2* (*DISC1* and *DISC2*) in schizophrenia (Millar et al., 2000b), *asparagine-linked glycosylation 9 homolog* (*ALG9*), formerly 'disrupted in bipolar disorder' (*DIBD1*) (Baysal et al., 2002), and *glutamate receptor, ionotrophic, AMPA 3* (*GRIA3*) in BPAD (Baysal et al., 2002; Gecz et al., 1999).

Appropriate methods to disentangle the genetics of complex illnesses have been debated. Two main genetic models suggest two different approaches (Jorde, 2000). The following two sections present these two major approaches, namely linkage analysis and linkage disequilibrium mapping, and the genetic models that underlie them.

1.1.1 Linkage analysis

The first model assumes rare variants each with an individual, detectable effect. Under this model, molecular cytogenetics, family linkage studies and functional candidate gene analysis would be capable of identifying regions and genes of interest, as they were successfully applied to single-gene illnesses (Botstein and Risch, 2003). Linkage is the tendency of DNA sequences at specific loci to be inherited together as a consequence of their physical proximity on a single chromosome, that is the recombination fraction between the two loci is less than 0.5. Linkage analysis attempts to detect linkage between genetic markers and the phenotype of interest in related individuals. Detection of linkage relies on the calculation of the overall likelihood of the observed pedigree on the alternative assumptions that the loci are linked, determined by a given recombination fraction, or not linked, where the recombination fraction is 0.5. The ratio of these two likelihoods indicates the odds of linkage, and the logarithm of the odds, the lod score, is used to determine whether the markers are linked: the higher the lod score, the

stronger the evidence for linkage (Morton, 1955). For a Mendelian character, with a 5% chance of error, the threshold for significance is 3.0 for a single marker, and 3.3 for a genome-wide analysis (Lander and Schork, 1994). Linkage analysis methods for complex disorders can be divided in parametric and non-parametric methods (Jorde, 2000). Parametric methods rely on the specification of the Mendelian mode of inheritance, the number of genes involved, the frequency of susceptibility alleles and their penetrance (frequency at which a genotype manifests itself in a given phenotype). Such methods are appropriate for the detection of genes for Mendelian disorders or genes of major effect. Study of the segregation of markers along a chromosome with the phenotype of interest requires large multigenerational families with several affected members. On the other hand, non-parametric methods do not rely on a specified model, ignore unaffected individuals and identify alleles or chromosome segments that are shared by affected people, in nuclear (sib pair) or extended families (Jorde, 2000).

In genome-wide linkage analysis of complex diseases, thresholds for significance have been suggested and are now widely used: 3.3-4.0 for an initial report, depending on the study design, and 1.2 for replication reports (Lander and Kruglyak, 1995). Linkage analyses of several hundreds of markers across the genome have so far generated mixed results due to initial reports of significant linkage of the illnesses to regions in the genome, but lack of confirmation by subsequent studies (Altmuller et al., 2001). Initial reports are likely to overestimate the size of effect while subsequent replication attempts are more likely to indicate the true size of effect (Lander and Kruglyak, 1995; Altmuller et al., 2001; Botstein and Risch, 2003). Some conflicting findings may reflect the complexity of the diseases and may be due to the involvement of several genes and environmental factors (Botstein and Risch, 2003). Two genetic models have been considered to explain the discrepancy between studies. Firstly, mutations in one gene (allelic heterogeneity) or in several genes (genetic heterogeneity) can generate very similar phenotypes, such as in retinitis pimentosa (Kennan et al., 2005). Extended pedigrees where a single major gene is segregating are therefore more suited to detection of linkage than smaller families and affected sib pairs. In the second model, several genes with a small effect interact to generate the observed phenotypes. To detect mutations

conferring low risk of disease, linkage analyses would require an unachievable number of families (Risch and Merikangas, 1996). Lack of replication would therefore be due to the lack of power of current studies (Blangero, 2004; Lander and Kruglyak, 1995). Imprecise diagnoses, in particular for psychiatric diseases, are also liabilities in linkage analysis (McGorry et al., 1995). This can be circumvented by carrying out analyses with several definitions of the phenotypes, such as in the case of affective disorders, by distinguishing unipolar depression and bipolar affective disorder (Blackwood et al., 1996). Failure to replicate can also result from population heterogeneity and statistical fluctuation (Lander and Kruglyak, 1995). Despite mixed results (Byerley, 1989; Risch and Botstein, 1996), several studies have reported highly significant linkage based on extended families and large sets of small families for several complex disorders, such as schizophrenia (SCZ) (St Clair et al., 1990; Stefansson et al., 2002), affective disorders (Blackwood et al., 1996), type 1 diabetes (Nistico et al., 1996) and inflammatory bowel disease (Rioux et al., 2000). Linkage regions are typically measured in megabases (Mb) and harbour tens or hundreds of genes. Such regions are usually narrowed further via linkage disequilibrium mapping (Botstein and Risch, 2003).

1.1.2 Linkage disequilibrium mapping

Linkage disequilibrium (LD) is the non-random association between alleles at different loci. The second approach to disentangle the genetics of complex diseases, LD mapping or association studies, builds on theoretical and experimental data that suggest a polygenic model for the genetic variation in complex traits. Because most complex diseases are also common, variants involved in their genetic aetiology may be multiple, common and potentially ancient sequences with small effect, acting in an additive manner (Risch and Merikangas, 1996). Under this model, genome-wide linkage analyses with a realistic sample size would fail to detect such small effect determinants (Risch and Merikangas, 1996). In contrast, association studies, which compare allele frequencies at a locus, microsatellite or Single Nucleotide Polymorphism (SNP), in patients and matched controls have the power to detect variants with modest effect in a sample of achievable size (Cardon and Bell, 2001; Risch, 2000; Risch and Merikangas, 1996).

The variant typed is either functional, and potentially disease-causing, or more likely, in LD with the causative variant. LD mapping is complicated by the uneven and unpredictable distribution of linkage disequilibrium in the human genome (Tiret et al., 2002). Some studies have shown that populations of different ethnic origin share a limited number of common haplotypes (series of alleles found at linked loci on a single chromosome) (Gabriel et al., 2002), while others have found that any population can be characterised by a small number of haplotypes and that haplotypes vary among populations (Kauppi et al., 2003). Population stratification is another confounding factor and can result in spurious association in case-control studies where controls do not match cases adequately (Clayton and McKeigue, 2001). Methods avail to test for stratification in the study set, as in family-based association studies relying on the transmission disequilibrium test for example (Ewens and Spielman, 1995). To avoid scoring markers at high, prohibitive densities required for whole-genome studies, a higher priority is given to variants within coding sequences (Botstein and Risch, 2003) and regulatory sequences (Mitchison, 2001). These markers are also chosen to capture the majority of the haplotypes in the population under scrutiny (Johnson et al., 2001; Goldstein et al., 2003). As haplotypes have been shown to be more informative than single markers, difference in haplotype frequencies between affected and unaffected samples are also tested for association with disease (Hirschhorn and Daly, 2005).

Association studies of variants in several positional or functional candidate genes for complex diseases have generated promising results (Lohmueller et al., 2003): *insulin* (Bell et al., 1984), *human leukocyte antigen* (Dorman et al., 1990) and *cytotoxic T lymphocyte associated-4* (*CTLA4*) (Nistico et al., 1996) for type 1 diabetes; *peroxisome proliferator-activated receptor-γ* (*PPARG*) for type II diabetes (Deeb et al., 1998; Altshuler et al., 2000); *phosphodiesterase 4D* (*PDE4D*) for stroke (Gretarsdottir et al., 2003); *lymphotoxin-alpha* (*LTA*) for myocardial infarction (Ozaki et al., 2002); *caspase recruitment domain family, member 15* (*CARD15*), and cytokines for inflammatory bowel disease (Rioux et al., 2000; Hugot et al., 2001; Ogura et al., 2001); *apolipoprotein E* (*APOE*) for Alzheimer's disease (Strittmatter and Roses, 1996) and *neuregulin* (*NRG1*) for schizophrenia (Stefansson et al., 2002).

Many genes could potentially be involved in the various symptoms observed in neuropsychiatric disorders, including psychoses such as affective disorders and schizophrenia, generating a very long list of candidate genes within linked regions. If several families from a stable population shared a common founder mutation, then typing markers across the linked region of each affected family may detect recombinants that define a common, smaller region. Markers in the shared region can then be tested for association using cases and controls from the same population to narrow the region further and identify positional candidate genes (Porteous et al., 2003).

1.2 Psychiatric disorders

Psychiatric disorders involve abnormal anxiety and obsession, mood, thought, cognition, personality, gender and sexuality. They include psychosis, schizophrenia, affective disorders, dementia, autism, attention-deficit hyperactivity disorder, Parkinson's disease, Huntington's disease and Alzheimer's disease. The lifetime risk for mental illness is 15%, and one to two thirds of serious cases are not treated (Bijl et al., 2003; Demyttenaere et al., 2004). Psychosis is characterised by derangement of personality and loss of contact with reality, causing deterioration of normal social functioning (American Psychiatry Association, 1994). It is the third cause of disability world-wide after quadriplegia and dementia and before paraplegia and blindness (Ustun, 1999). To date, 10% of the total expenditure of the UK National Health Service is used to treat mental illness, and the medical, social and economic costs supported by the 50 million strong population are estimated at £32 billion per annum (Simon, 2003). Research into psychiatric disorders includes four strategies: i) detection of structural abnormalities in the brain, ii) assessment of the effect of drugs on target receptors and on physiology of the brain, iii) linkage and association studies of functional or positional candidates genes, and iv) study of epigenetic factors in disease causation (Inoue and Lupski, 2003). Recent progress in the understanding of Alzheimer's disease (AD) aetiology, a neurodegenerative disorder with psychiatric symptoms, such as depression, illustrates the usefulness of combining cytogenetic, linkage and association analyses (Munoz and Feldman, 2000). A great proportion of individuals with Down's syndrome, which is caused by the presence of a third

chromosome 21, suffer from AD. This observation led to the identification of a mutation in the *amyloid precursor* gene (*APP*) (van Leeuwen et al., 1998). The *presenilin 1* (*PS1*) gene was identified by linkage analysis of families with AD, while the *presenilin 2* (*PS2*) gene was identified by homology searches with *PS1* (Sandbrink et al., 1996). Association studies identified the *apolipoprotein E* (*ApoE*) gene in a region on chromosome 12 previously detected by linkage analysis (Roses, 1994). Biological studies subsequently showed that *APP* plays a role in learning and that its activity is regulated by presenilins (Munoz and Feldman, 2000).

Affective disorders and schizophrenia (SCZ) are common and debilitating diseases which respectively involve disturbed mood and thought, varying in severity among patients. Both illnesses often appear in early-adult life and occur throughout the entire lifespan (Gelder et al., 2000). The lifetime risk from each of BPAD and SCZ is estimated at 1% (Weissman et al., 1988). Because these psychoses can be particularly disabling and are frequently chronic and relapsing, caring for patients represents a considerable load for the patient's family, and for the social and health services (Gelder et al., 2000). The annual cost of depression and SCZ in the United States has been estimated at $83 billion and $32.5 billion, respectively (Rice, 1999; Greenberg et al., 2003). The World Health Organisation has predicted that unipolar depression alone will become the disease placing the heaviest burden of disability and mortality on the population of developed countries by 2020 (Murray and Lopez, 1997).

1.2.1 Clinical features of affective disorders and schizophrenia

1.2.1.1 Affective disorders

The earliest description of depression was found in ancient Egyptian medical papyri, dating approximately from 2000 BC (Scholl, 2002). The concept of bipolar manic-depressive illness, now termed bipolar affective disorder (BPAD), was provided by Arataeus, a Roman physician in 2nd century AD (Adams, 1856). It was later described as 'folie circulaire', that is 'cyclical insanity', by Falret in 1854 who realised that cycles of depression and elation were different from simple depression, and were "generally found in similar form in ascendants and descendants" (Falret,

1854; Sedler, 1983), and by Baillarger in the same year as 'folie à double forme', meaning 'insanity of double form' (Baillarger, 1854). Further distinction of BPAD from other mental illnesses, including schizophrenia, was defined by Emil Kraepelin in 1899, who believed that depression, then called 'melancholia', and symptoms now defining BPAD were exogenous and thus treatable, while schizophrenia was caused by organic changes in the brain, hence endogenous and incurable (Kraepelin, 1899).

At present, psychiatric diagnostics are based on profiles of symptoms, such as behavioural indices and communication patterns, which were standardised to allow for reproducibility of diagnoses between psychiatrists (American Psychiatry Association, 1994). Affective disorders can have a uniquely depressive (unipolar) or manic-depressive (bipolar) course, or occasionally uniquely manic. A major depressive episode is defined as a period of at least two weeks during which there is a onset or worsening of depressed mood or dysphoria, that is characterised by loss of interest or enjoyment of nearly all activities, together with four of the following symptoms: changes in appetite, weight, sleep and psychomotor activity; decreased energy; feelings of worthlessness or guilt; difficulty in thinking, concentrating, or making decisions; or recurrent thoughts of death or suicidal ideation, plans, or attempts. These symptoms must cause distress or impairment in social, occupational, or other important area of functioning (American Psychiatry Association, 1994). Unipolar disorder (UP) is characterised by one or more episodes of depression without a history of manic, mixed or hypomanic episodes (described below). UP is often recurrent and is one of the most common psychosis, as it affects between 10% and 25% of women and 5.5% to 12% of men (American Psychiatry Association, 1994). Among patients suffering from severe depressive disorder 10-15% eventually commit suicide, and co-morbid panic disorder is common in women, while co-morbid substance abuse is more common in men (American Psychiatry Association, 1994; Gelder et al., 2000). The rate of suicide is also elevated in schizophrenics, 10% (Caldwell and Gottesman, 1990), while 30% admit to have attempted suicide at least once (Leucht et al., 2003). Depression is also observed in 20% to 45% of patients with Parkinson's disease and is also common in patients with multiple sclerosis (Rickards, 2005). Mania is characterised by abnormally and persistently elevated mood and/or irritability accompanied by at least three of the following

symptoms: over-inflated self-esteem, decreased need for sleep or insomnia, racing thoughts, physical agitation, increased talkativeness, distractibility, increased goal-directed activity such as shopping, hypersexuality, and excessive involvement in risky behaviour or activities. Hypomania is a less severe form of mania (American Psychiatry Association, 1994).

BPAD is characterised by severe swings of mood to depression or mania, as well as mixed manic and depressive symptoms. These symptoms may be accompanied by hallucinations and delusions (described below). Episodes of depression and mania may be separated by periods of wellness, free of symptoms, during which the patient is able to function normally both at work and in social affairs. Rapid cycling BPAD is characterised by one or more episodes of illness within a 12-month period. Bipolar affective disorders are classified into two categories. BPAD I is defined by the occurrence over the course of the individual's life of at least one manic or mixed state episode which is usually accompanied by episodes of depression. BPAD II is defined by the occurrence over the course of the individual's life of at least one episode of depression which must be accompanied by at least one hypomanic episode (American Psychiatry Association, 1994).

In the brain, the prefrontal cortex is involved in regulation of mood and responses to social stimuli. In patients with UP or BPAD, this region is smaller and shows decreased activity (Drevets et al., 1997). Treatment of affective disorder uses mood stabilisers to prevent or mitigate manic or depressive episodes, such as antidepressants, lithium salts and anticonvulsants such as valproic acid and carbamazepine (Harwood and Agam, 2003). Lithium is thought to substitute for magnesium ions in the active site of inositol monophosphatase, which dephosphorylates inositol monophasphate into inositol and phosphate, thereby reducing its activity and affecting neurotransmitters abundance (Williams et al., 2004c). Valproic acid has been prescribed to treat mania and seems to target intracellular inositol too (Williams et al., 2002; Harwood and Agam, 2003; Shaltiel et al., 2004). The episodic nature of BPAD complicates treatment as patients naturally experience periods of remission. Furthermore medication doses must be monitored closely as mood stabilisers often have secondary effects which include dry mouth, blurred vision, constipation, sedation, weight gain and changes in sexual function, and can be cardiotoxic at high level (Rickards, 2005).

1.2.1.2 Schizophrenia

As for depression, description of symptoms akin to that of SCZ can be found in the ancient Egyptian medical papyri (Scholl, 2002). It is also found in Hindu text from 1400 BC, containing hymns and incantations to correct an imbalance of the eight health-controlling elements thought to result in madness (Max-Muller, 1998). A Chinese text from 1000 BC also describes insanity (Ti and Suwen, 1966). At that time, psychotic behaviours were often explained by supernatural possession, including demons and gods (Max-Muller, 1998). This view was abandoned few centuries later by the Greeks during the fifth and fourth centuries BC who tempted to rationalise the relationship between mind and body, especially Plato (Jowett, 1898), and that between mental disorder and the brain, notably Hippocrates (Blits, 1999), as well as the Roman Arataeus in 2nd century AD (Adams, 1856). Symptoms of SCZ, were then better described in the late eighteenth century by Pinel, one of the founders of modern psychiatry (Pinel, 1798), and later termed 'démence précoce', that is 'early dementia', by Morel, an assistant of Falret who described 'cyclical insanity', to refer to a mental and emotional deterioration which begins at the time of puberty, in contrast to senile dementia, ending in mental illness, and which involves heredity as well as abuse of alcohol and drugs (Morel, 1860). The term 'démence précose' was translated into 'dementia preacox' by Kraepelin in 1899 to distinguish between BPAD and SCZ, by noticing that these illnesses were best defined by specific patterns of symptoms rather than by any particular symptom, and that the course of BPAD was intermittent with symptom-free periods separating manic or depressive episodes, while the course of SCZ was progressive and irreversible (Kraepelin, 1899). This early clinical description of SCZ synthesised existing concepts into a workable model which comprised three classes: i) catatonia, in which motor activities are disrupted, either excessively active or inhibited, ii) hebephrenia, characterised by inappropriate emotional reactions and behaviour, and iii) paranoia, characterised by delusions or grandeur and persecution (Kraepelin, 1899). A couple of decades later, Bleuler believed that 'dementia praecox' did not fit symptoms of current patients, some of which did not have dementia nor showed deterioration early in life, and opposed the view that so called 'dementia praecox' was incurable. In 1911, Bleuler coined the term 'schizophrenia' from the Greek for split, 'schizo', and

mind, 'phrene', to characterise the disharmonious state of mind, and introduced the notions of autism, the loss of contact with reality, and ambivalence, the coexistence of mutually exclusive contradictions in the mind (Bleuler, 1911).

At present, SCZ is characterised by symptoms such as cognitive dysfunction, disordered thinking, affective changes, and catatonia, which are classified as positive, when they are additional to normal experience and behaviour, and negative, to express the lack or decline in normal experience or behaviour. Positive symptoms include hallucinations, delusions, and paranoia. Hallucinations are characterised by hearing, seeing, or sensing the presence of stimuli that do not exist. Delusions are false personal beliefs that are not subject to reason or contradictory evidence and are not explained by a person's cultural background. Negative symptoms comprise disorganised thought and speech, flattened affect, apathy, inability to experience pleasure, impaired attention and poor social skills including social withdrawal. To be diagnosed as schizophrenic an individual must display over a significant portion of time during a 1-month period a combination of two or more characteristic symptoms, including hallucinations, delusions, frequently derailed or incoherent speech, disorganised or catatonic behaviour. These signs of disturbance must persist for at least six months and lead to social or occupational dysfunction in one or more of the following areas: work, interpersonal relations or self-care (American Psychiatry Association, 1994). SCZ affects emotions, but is distinguished from mood disorders where affective symptoms are primary. Similarly, SCZ and dementia are distinguished on the basis that cognition is most severely impaired in dementia (American Psychiatry Association, 1994).

Five sub-classifications of SCZ exist: i) catatonic, ii) disorganised, where thought disorder and flat or inappropriate mood co-occur, iii) paranoid, characterised by hallucinations and delusions in the absence of disorganised thought, behaviour or mood, iv) residual, where intensity of positive symptoms is low, and v) undifferentiated, where psychotic symptoms exist but criteria for catatonic, disorganised and paranoid classes are not met (American Psychiatry Association, 1994). These symptoms of SCZ are often accompanied, and may be caused by, specific neurocognitive deficits characterised by impairment or reduction in basic psychological functions such as memory, attention, problem solving and executive

function (Gelder et al., 2000). Structural neuroimaging and neuropathological studies have suggested that SCZ may result from abnormal neurodevelopment, especially in the second trimester of foetal life (Woods, 1998). Structural abnormalities observed in brain of SCZ patients, such as decreased volume of overall white matter and prefrontal cortex, were correlated with severity of negative symptoms (Sanfilipo et al., 2000; Wible et al., 2001). Functional differences in activity in the brain of schizophrenics have been reported in the frontal lobes, hippocampus, and temporal lobes (Green et al., 2001). Many drugs used to treat SCZ are dopamine receptor antagonists and typically have secondary effects in schizophrenics such as rigidity, bradykinesia, dyskinesia (Malhotra et al., 2004). Drugs used to treat SCZ are generally referred to as antipsychotic and may be classified as 'typical', such as chloropramazine and haloperidol, and 'atypical', such as clozapine, risperidone, olanzapine, because the latter avoid side effects associated with prolonged treatment with 'typical' dopamine antagonists. However, atypical antipsychotics appear to have metabolic side effects, such as weight gain, hyperglycemia and hypertriglyceridemia (Leucht et al., 2003).

1.2.2 Aetiology of affective disorders and schizophrenia

There is evidence from twin, adoption and family studies for non-genetic as well as genetic effects in affective disorders and SCZ (Bertelsen et al., 1977; Baron, 1997; Merikangas and Risch, 2003). For example, heavy use to stimulants, such as methamphetamine, and hallucinogenic drugs, such as LSD, ketamine, or phencyclidine, so-called 'angel dust', can result in syndromes similar to those observed in SCZ (Joyce et al., 1993; Goodman, 2002). Moreover, there are structural similarities between the hallucinogen LSD and medical target serotonin (Johnson et al., 1991), in keeping with the reduced concentrations of serotonin in the brain of schizophrenics after treatment by medical drugs, such as typical antipsychotic reserpine (Stahl and Wets, 1988). Other environmental factors include stressful life events, latitude, urban birth, household crowding, having older siblings and famine during pregnancy (Huttunen and Niskanen, 1978; Day et al., 1987; Yolken and Torrey, 1995; Susser et al., 1996; Sundquist et al., 2004). Some of these correlations

might be explained by certain infections, perinatal brain damage, stillbirths, prematurity and nutritional deficiencies (Torrey et al., 1997). For example, an excess of 5 to 8% of SCZ birth has been observed in the winter-spring months, in particular in individuals with no SCZ family history, and may involve viral infections during early pregnancy leading to disruption of neurodevelopmental processes by viral toxic substances (Torrey et al., 1997; Davies et al., 2003). Other cases may involve interactions between environmental influences and genetic predisposition (Risch and Merikangas, 1996).

Though little is understood about the underlying causes of major affective disorders and SCZ, strong support for a genetic aetiology has arisen from family, twin and adoption studies (Bertelsen et al., 1977; Baron, 1997; Merikangas and Risch, 2003). The estimates of heritability for BPAD and SCZ vary but are as high as 80% (Hoeffer and Pollin, 1970; Owen and Cardno, 1999). The risk to a first-degree relative of an individual affected by BPAD or SCZ is approximately ten times that to a member of the general population (Merikangas and Risch, 2003). The age of first diagnosis in early adulthood suggests a time for expression of symptoms that is biologically and genetically influenced, and a potential susceptibility to life changes and exposures that depend on age (Merikangas and Risch, 2003). Genetic analyses of psychoses are confounded by the current classification based on diagnoses, which are reported by patients and hence often imprecise, by the lack of reliable biological and genetic markers, anticipation and parent-of-origin effect (Gelder et al., 2000; Merikangas and Risch, 2003).

Anticipation is the phenomenon by which the age of onset of the disorder is reduced and / or the severity of the symptoms increases in subsequent generations. It can be due to observation bias (Bassett and Husted, 1997; Husted et al., 1998) or environmental factors, but recent genetic studies of neurodegenerative diseases such as muscular dystrophy, fragile X syndrome and spinal muscular atrophy have been shown to be caused by trinucleotide repeat expansion (Teisberg, 1995; McInnis, 1996; Bayless et al., 1998; Kenneson et al., 2001). Anticipation has been observed in several neuropsychiatric diseases, including BPAD and SCZ, and tentatively associated with trinucleotide expansions (McInnis et al., 1993; Stober et al., 1995;

O'Donovan et al., 1996; Mendlewicz et al., 1997; Lindblad et al., 1998; McInnis et al., 1999b; O'Donovan et al., 2003a; Swift-Scanlan et al., 2005) but these findings have not been replicated or specific examples of trinucleotide repeat expansion demonstrated (O'Donovan et al., 1996; Craddock et al., 1997; Visscher et al., 2001; Lange and McInnis, 2002; O'Donovan et al., 2003a; Tsutsumi et al., 2004).

The 'parent-of-origin' effect refers to the higher number of affected mothers and maternal relatives than affected fathers, and paternal relatives, or vice-versa, and has been observed in bipolar patients (Winokur and Reich, 1970). It might reflect parent-of-origin specific imprinting or mitochondrial inheritance (McMahon et al., 1995; Stober et al., 1998a). For example, McMahon et al. (1995) reported significantly higher frequency of affected mothers, increased risk of illness for maternal relatives and offspring, but no paternal transmission of BPAD was detected. The authors suggested that these observations may indicate a potential role for mtDNA and imprinted DNA in affective disorders. Also, linkage to 18p11 is limited to families where the disease is transmitted paternally (Stine et al., 1995; Escamilla et al., 1999; McInnes et al., 2001) while maternal effect on 6q16.3-22.1 is suggested by affected siblings sharing the maternal chromosome more often than the paternal chromosome (Schulze et al., 2004). Association of a mtDNA polymorphism and BPAD has also been suggested and might be a risk factor with small effect (Kato et al., 2001; Kato and Kato, 2000; McMahon et al., 2000).

Psychosis is currently thought to be the result of a combination of polygenic inheritance, genetic and allelic heterogeneity, and variable exposures to environmental factors, including stress (Merikangas and Risch, 2003). The identification of susceptibility genes is required to unravel interactions among genes, between genes and environmental factor, and the role of epigenetics factors (Harrison and Weinberger, 2005). The next sections provide details on the genetic determinants of affective disorders and schizophrenia.

1.2.3 Genetic basis of affective disorders

1.2.3.1 Positional candidate genes

Genes implicated in genetic diseases associated with BPAD are putative candidates genes for affective disorders (Kato, 2001b). Some single-locus diseases have phenotypes similar to that found in common diseases. Alterations in the genes for such Mendelian diseases may be implicated in the aetiology of common disorders. For example, Wolfram syndrome, characterised by diabetes mellitus, optic atrophy, deafness and depression is an autosomal recessive disease caused by mutation in the *WFS1* gene at 4p16 (Inoue et al., 1998). Non-affected, heterozygous carriers of this gene constitute 1% of the population and have a 26-fold increased risk of psychiatric disorder, suggesting that *WFS1* may therefore be implicated in psychiatric diseases, although this role has yet not been established (Swift and Swift, 2000).

1.2.3.1.1 11q 23 and ALG9

Other genes involved in affective disorders have been identified using a cytogenetic approach. A reciprocal chromosomal translocation t(9;11)(p24;q23) has been identified in patients with UP and BPAD, disrupting a gene, then called *disrupted in bipolar disorder* (*DIBD1*) and later renamed *asparagine-linked glycosylation 9 homolog* (*ALG9*) at 11q23 (Baysal et al., 2002). Regions within 11q21 to 11q25 have been linked to SCZ (St Clair et al., 1990), schizoaffective disorder (Holland and Gosden, 1990) and Tourette syndrome (Simonic et al., 2001), a neurologic disorder characterised by motor and vocal tics and associated with behavioural abnormalities (American Psychiatry Association, 1994). The *ALG9* gene encodes a mannosyltransferase which is involved in the N-glycosylation pathway by adding mannose residues to lipid-linked oligosaccharides (Burda et al., 1996). ALG9 has been implicated in congenital disorder of glycosylation (CDG) (Frank et al., 2004), which is characterised by failure to thrive, axial hypotonia, cerebellar dysfunction, seizures, mental retardation and coagulopathy (Krasnewich and Gahl, 1997). The role of ALG9 in BPAD however remains a open question (Baysal et al., 2002; Frank et al., 2004).

1.2.3.1.2 1q43 and DISC1/DISC2

Linkage between a t(1; 11)(q43;q21) balanced reciprocal translocation and major psychiatric illness, including UP, BPAD and SCZ was found in a Scottish family (St Clair et al., 1990; Muir et al., 1995; Blackwood et al., 2001). The t(1;11)(q43;q21) breakpoint directly disrupts two overlapping genes of unknown function, *DISC1* and *DISC2*, which are located on chromosome 1 and expressed in the brain. DISC1 is predicted to have a globular head, with no similarity to any known proteins, and a tail with a putative ability to interact with other proteins (Ma et al., 2002; Taylor et al., 2003). DISC1 has been shown to interact with NUDEL, a cytoskeletal protein expressed in the cortex and involved in neuronal migration (Ozeki et al., 2003). Interactions between DISC1 and NUDEL may play an important role in early brain development (Brandon et al., 2004). Other potential interactors with known neuronal function have been predicted (Millar et al., 2003), such as FEZ1 which is implicated in axon guidance, cell adhesion and neurogenesis (Miyoshi et al., 2003). DISC1 is predominantly associated with mitochondria, but is also observed in the cytoskeleton and the nucleus (Ozeki et al., 2003; James et al., 2004b). These findings suggest that DISC1 may serve as a 'scaffold' protein and be involved in several neuronal processes, such as signal transduction, cytoskeletal alteration, neuronal migration and neurite outgrowth (Millar et al., 2004). *DISC2* seems to be a non-coding RNA gene antisense to *DISC1* and might therefore regulate DISC1 expression (Millar et al., 2000b). Upstream of *DISC1* lies the *Translin-associated factor X* (*TRAX*) gene which undergoes intergenic splicing with *DISC1* (Millar et al., 2000a). TRAX binds to and is stabilised by Translin, which migrates to neuronal dendrites upon neuronal activity (Kobayashi et al., 1998; Yang and Hecht, 2004). In the brain of mutant mice deficient for *Translin*, and thus for *TRAX*, expression of genes encoding neurotransmitter receptors and ion channels is decreased (Chennathukuzhi et al., 2003). The disruption of *DISC1* and *DISC2*, and additional genetic and environmental factors may be sufficient to explain psychosis observed in carriers of the translocation (Millar et al., 2003). Recently, SNPs and haplotypes in the *TRAX/DISC1* region have been associated with BPAD, schizoaffective disorder and SCZ (Hennah et al., 2003; Hodgkinson et al., 2004; Thomson et al., 2005).

1.2.3.1.3 Other regions

Positional candidate genes are generally identified due to their location in chromosomal regions linked to the disease. Several candidate regions for BPAD have been identified by two or more linkage studies: 1q21-42, 4p16, 4q, 6p, 10q21-26, 11p15.5, 12q23-24, 13q11-32, 15q, 18p11, 21q and 22q11-12 (Potash and DePaulo, Jr., 2000; Baron, 1997; Owen and Cardno, 1999; Sklar, 2002). Linkage of the chromosome 4p15-16 region to UP and BPAD has first been reported by our laboratory (Blackwood et al., 1996). Subsequent studies have found linkage to this region for BPAD (Ewald et al., 1998; Detera-Wadleigh et al., 1999) and schizoaffective disorder (Asherson et al., 1998). Association between markers at 4p15-16 and BPAD and SCZ, and increased allele sharing in schizophrenics has also been reported (Als et al., 2004; Williams et al., 1999). Positional cloning in this region is being carried out in the laboratory to uncover putative susceptibility genes for these diseases. Continuing work aims to assess the roles of these genes in the aetiology of these illnesses (chapter 6).

1.2.3.2 Functional candidate genes

The anti-depressive effect of monoamine reuptake inhibitors and the anti-manic effects of monoamine receptor antagonists suggest that alteration of the monoaminergic neurotransmission is involved in the pathophysiology of affective disorders. Genes implicated in neurotransmitter pathways, such as the dopaminergic, serotonergic and glutamatergic systems are therefore potential candidate genes (Merikangas and Risch, 2003). These genes either encode the neurotransmitters or their receptors, or are involved in the uptake, synthesis, transport, reuptake and degradation of neurotransmitters. Also, response to treatment, in particular with antidepressant drugs, has been suggested to be partly heritable (Pare and Mack, 1971; O'Reilly et al., 1994; Franchini et al., 1998).

1.2.3.2.1 Serotonergic transmission

Serotonin governs several behavioural aspects such as sleep and mood (Malhotra et al., 2004). Association of the gene encoding the rate-limiting enzyme for serotonin synthesis in the brain, tryptophan hydroxylase 2 (TPH2), has been associated with UP (Zhang et al., 2005). Seven serotonin receptors have been

identified: 5HT-1 to 5HT-7, with several subtypes for 5HT-2 and 5HT-3 (Barnes and Sharp, 1999). Variants in three receptors have been reported to be associated with BPAD: 5HTR2A, 5HTR2C and 5HTR6 (Fanous et al., 2004; Malhotra et al., 2004) and response to antidepressant (Minov et al., 2001; Cusin et al., 2002b). The *serotonin transporter* (*SLC6A4*) gene mediates serotonergic neurotransmission by controlling magnitude and duration of the serotonin signal (Lesch and Mossner, 1998) and is an important gene involved in neurodevelopment, as well as functioning and plasticity of the brain (Catalano, 2001).

Two polymorphisms in *SLC6A4* have also been associated with affective disorders in some but not all studies. First, a 44bp insertion/deletion in the promoter region (Heils et al., 1996), which regulates the expression of the gene (Heils et al., 1996; Mortensen et al., 1999) and affects the abundance (Heinz et al., 2000) and function (Hariri et al., 2002) of SLC6A4. This variant has also been reported to be associated with BPAD (Rotondo et al., 2002; Collier et al., 1996) and anxiety- and depressive-related personality traits of neuroticism (Lesch et al., 1996). Second, a variable number of tandem repeats (VNTR) in the second intron (Lesch et al., 1994) upregulates the expression of the *SLC6A4* gene (Fiskerstrand et al., 1999; Hoefgen et al., 2005; Lasky-Su et al., 2005) in a manner dependent on the number and structure of the VNTR (Lovejoy et al., 2003). This variant has been associated with UP (Ogilvie et al., 1996) and seems to act in combination with the insertion/deletion variant in the promoter (Hranilovic et al., 2004). Stronger evidence for the implication in anxiety of the latter polymorphism was obtained by a study of the association of an endophenotype for anxiety and fear behaviour: the response of the amygdala to fearful stimuli, measured by functional MRI. Differences between two haplotypes in neuronal activity in amygdala were 10-fold stronger than in previous studies when the endophenotype was not considered (van Rijn et al., 2005). The role of SLC6A4 in affective disorders is further supported by association of the promoter variant with response to selective serotonin reuptake inhibitors used to treat major depression (Smeraldi et al., 1998; Pollock et al., 1998; Zanardi et al., 2000; Kim et al., 2000; Minov et al., 2001; Yoshida et al., 2002; Serretti et al., 2004). Polymorphisms in the *SLC6A4* gene, including variants in its promoter have also been associated with Alzheimer's disease (Li et al., 1997a), autism in some studies (Conroy et al., 2004; Kim et al., 2002), obsessive-compulsive disorder (Ozaki et al., 2003), attention-deficit and hyperactivity disorder (ADHD) (Curran et al., 2005).

1.2.3.2.2 Dopaminergic transmission

Dopamine systems are involved in a wide spectrum of neurological and psychiatric disorders, such as Parkinson's disease, schizophrenia, drug abuse, affective disorders, Tourette's syndrome and ADHD (Bannon et al., 2001; Diehl and Gershon, 1992; Emilien et al., 1999). Tyrosine hydroxylase (TH) is a rate-limiting enzyme implicated in the metabolism of dopamine, adrenaline and noradrenaline (Kobayashi et al., 1995; Zhou et al., 1995). The *TH* gene lies on 11p15, which is linked to affective disorders (Sklar, 2002). Association studies of variants in or around the *TH* gene and affective disorders have reported positive results for BPAD (Meloni et al., 1995) and UP (Souery et al., 2001; Serretti et al., 2000b), as well as many negative results (Burgert et al., 1998; Serretti et al., 2000b; Souery et al., 1999; Furlong et al., 1999b; Ishiguro et al., 1998; McQuillin et al., 1999). These results may be confounded by epigenetic factors since the 11p15 region harbours a cluster of imprinted genes and alteration of the expression of the *TH* gene might be due to abnormal imprinting rather than mutation (Aranyi et al., 2002; Kim et al., 2003).

Five dopamine receptors have been identified and are classified in two groups: i) D1-like receptors, namely dopamine receptors D1 (DRD1) and D5 (DRD5), and ii) D2-like receptors, namely DRD2, DRD3 and DRD4. All are transmembrane proteins of the guanine nucleotide-binding (G-protein) receptor family. Dopamine receptors mediate stimulation of adenyl cyclase and intracellular accumulation of cAMP (Sakolsky and Ashby, 2001). D1-class receptors are required for induction of transcription factors of the nerve growth family (Young et al., 1991) and receptors in the two classes interact to modulate neural response (Walters et al., 1987). For example, acute behavioural effects mediated by D2-class receptors require the simultaneous activation of D1-class receptors (Capper-Loup et al., 2002).

DRD1 modulates N-methyl-D-aspartate (NMDA) receptors functions by direct interactions (Lee et al., 2002). A haplotype at two SNPs in the upstream region of the *DRD1* gene has been associated with BPAD (Ni et al., 2002) and one of these SNPs, in the 5'UTR, has been associated with BPAD I (Severino et al., 2005). No evidence for the functional significance of these findings has yet been found, and these mutations may therefore be in LD with disease causing variants (Severino et al., 2005). DRD5 has a higher affinity for dopamine that DRD1 (Grandy et al., 1991)

and binds directly to the a sub-unit of the γ–butyric acid (GABA) type A ($GABA_A$) receptor to regulate the strength of the synapse (Liu et al., 2000). *DRD5* reside in the 4p16 region which has been linked to BPAD (Visscher et al., 1999; Polymeropoulos and Schaffer, 1996; Blackwood et al., 1996; Ewald et al., 1998; Detera-Wadleigh et al., 1999) and schizoaffective disorder (Asherson et al., 1998). A protective locus has also been mapped to this region (Ginns et al., 1998). Markers in or around the *DRD5* gene do not seem to be associated with affective disorders (Kirov et al., 1999; Asherson et al., 1998).

The *DRD2* gene is located on 11q23, which has been linked to affective disorder (Nanko et al., 1994). DRD2 is located on post-synaptic dopaminergic neurons involved in reward-mediating pathways, locomotion, hormone production and drug abuse, and is a target of drugs to treat psychoses (Neville et al., 2004). A microsatellite in intron 2 of the gene has been associated with BPAD but not UP (Massat et al., 2002). Several studies have however failed to report linkage or association to affective disorders (Souery et al., 1996; Bocchetta et al., 1999; Kirov et al., 1999; Serretti et al., 2000b; Holmes et al., 2002). The *DRD3* gene does not seem to be associated with BPAD (Mitchell et al., 1993; Shaikh et al., 1993; Gomez-Casero et al., 1996; Souery et al., 1996; Kirov et al., 1999; Elvidge et al., 2001). Nonetheless, DRD3 is localised in areas of the brain implicated in cognitive, emotional and endocrine functions, and DRD3 mRNA levels are reduced in BPAD patients, independent of severity of symptoms and medication history (Vogel et al., 2004). Its expression is modulated by BDNF during development and adult life (Guillin et al., 2001). Moreover, DRD2 and DRD3 are both implicated in some effects of antipsychotic used to treat Parkinson's disease (Sokoloff et al., 1990).

The role of DRD4 in the aetiology of mood disorders is suggested by animal (Rubinstein et al., 1997), pharmacological (Tang et al., 1997; Drevets et al., 2001) and postmortem studies (Bowden et al., 1997). A SNP located 1.2 Kb upstream of the first exon of the *DRD4* gene has also been associated with better handling of conflict in reaction time experiments, which correlated with brain activity (Fan et al., 2003a; Fossella et al., 2002). Associations studies of the 48-bp repeat in the upstream region of the *DRD4* gene have generated positive results for UP (Manki et al., 1996) and BPAD (Lim et al., 1994; Muglia et al., 2002), as well as negative results (De

Bruyn et al., 1994; Serretti et al., 1998; Bocchetta et al., 1999; Li et al., 1999; Serretti et al., 1999; Cusin et al., 2002a; Serretti et al., 2002). However, studies that failed to replicate reported positive results may have lacked sufficient power to detect allelic association with disease (Lohmueller et al., 2003). Several coding variants have been identified, including a 48-pb sequence, in exon 3, coding for 16 amino acids located in the third intracellular loop and which exists in forms containing two to eleven repeats with allele frequencies varying widely across populations (Van Tol et al., 1992; Chang et al., 1996). Association of the seven-repeat allele has also been reported with childhood inattention and dysphoria (Levitan et al., 2004b), binge eating and subsequent weight gain in women with seasonal affective disorder (Levitan et al., 2004a).

The *dopamine transporter* (*SLC6A3*) gene at 5p15.3 is a plasma membrane transport protein that mediates dopamine neurotransmission temporally and spatially by rapidly re-accumulating dopamine after it has been released in the extracellular space (Bannon et al., 2001). Some data suggest that the number of repeats in a 40 bp VNTR in the 3' end of the *SLC6A3* gene is involved in regulation of the gene in dopaminergic neurones and is correlated with SLC6A3 ligand binding in vivo (Michelhaugh et al., 2001). Variants in the promoter region and in introns of the *SLC6A3* gene also result in a 2-fold difference in gene expression (Greenwood and Kelsoe, 2003). SNPs and haplotypes at these loci have been associated with BPAD (Greenwood et al., 2001; Hawi et al., 2003; Keikhaee et al., 2005).

1.2.3.2.3 Neurotransmitter catabolism

Proteins involved in the catabolism of monoamine have also been tested for association with affective disorders. Catecol-O-methyltransferase (COMT) is expressed in peripheral tissues and in the brain, primarily in neurones, and is an enzyme implicated in the catabolism of catecolamines, including dopamine, adrenaline, noradrenaline as well as drugs used to treat hypertension, asthma and Parkinson's disease (Mannisto and Kaakkola, 1999; Matsumoto et al., 2003a; Matsumoto et al., 2003b; Harrison and Weinberger, 2005). The *COMT* gene at 22q11, a region linked to BPAD (Kelsoe et al., 2001), has been associated with BPAD (Shifman et al., 2004). Furthermore, the low activity allele of the *COMT* gene

causes higher synaptic concentration of catecholamine and has been associated with rapid cycling BPAD (Papolos et al., 1998).

Other enzymes may also be implicated, e.g. monoamine oxydases. Monoamine oxidases break down excess dopamine and serotonin and are present at both pre-synaptic terminal and post-synaptic cells. *Monoamine oxidase A* (*MAOA*) at Xp11.23 encodes an enzyme that catabolises monoamines on the outer membrane of mitochondria. MAOA deficiency has been associated with abnormal amine metabolism, mental retardation and impulsive aggressive behaviour in human (Brunner et al., 1993) and mouse (Cases et al., 1995). Polymorphisms in the *MAOA* gene, including a microsatellite in intron 2, has been associated with BPAD, with different alleles being associated in different populations (Lim et al., 1995; Kawada et al., 1995; Rubinsztein et al., 1996), although some studies could not replicate these findings (Craddock et al., 1995; Nothen et al., 1995). A functional microsatellite has been identified 1.2 Kb upstream of the coding sequence, with short alleles causing lower gene expression (Sabol et al., 1998; Deckert et al., 1999). This polymorphism has been associated with affective disorders (Lin et al., 2000; Schulze et al., 2000) but not by all studies (Kunugi et al., 1999; Furlong et al., 1999a; Syagailo et al., 2001; Huang et al., 2004). In addition, in males with a history of child abuse, low-expression alleles have been associated with impulsivity (Huang et al., 2004), while long, high-expression alleles have been associated with panic disorder in females (Deckert et al., 1999). Association studies of other variants have generated negative studies (Muramatsu et al., 1997; Sasaki et al., 1998).

Polymorphisms in genes encoding monooxygenases of the cytochrome P450 superfamily, such as *CYP2D6*, have also been associated to response to antidepressant treatment (Fukuda et al., 1999; Yoon et al., 2000; Malhotra et al., 2004; Johnson et al., 2005) and genotypes at loci in such genes, including *CYP2D6* and *CYP2C19*, have been proposed to establish dose of antidepressant (Kirchheiner et al., 2001; Athanasiou et al., 2002; Guzey and Spigset, 2004; James et al., 2004a). Another oxidase, D-amino acid oxidase (DAO) that oxidises D-serine, an activator of the NMDA-type glutamate receptor (Heresco-Levy et al., 1996; Mothet et al., 2000), may be involved in affective disorders since the *DAO* gene lies in 12q34 which may be linked to BPAD (Ewald et al., 2002; Shink et al., 2005) but the only association

study of *DAO* with BPAD generated negative results (Schumacher et al., 2004). However, DOA physically interacts with and is activated by DAO activator (DAOA), formerly G72 (Mothet et al., 2000; Chumakov et al., 2002). Interestingly, the *DAOA* gene lies in 13q32-33, linked to BPAD (Badner and Gershon, 2002), and has been associated with BPAD (Hattori et al., 2003; Schumacher et al., 2004).

1.2.3.2.4 Effects of mood stabilisers

Mood stabilisers such as lithium, valproic acid and carbamazepin, which are used to treat BPAD, interact with two signalling pathways, the first involving intracellular inositol, and the second glycogen kinase synthase 3. These three drugs suppress inositol signalling via an inositol depletion mechanism involving propyl oligopepdidase, which regulates inositol metabolism, and sodium myo-inositol transport (Harwood and Agam, 2003). They also alter neuronal morphology by inhibiting the collapse of sensory neurone growth cones and increasing growth cone area (Williams et al., 2004c). Lithium has been shown to reduce expression and activity of the inositol monophosphate 2 (IMPA2) (Seelan et al., 2004). Very recently, a SNP in the promoter of the *IMPA2* gene has been associated with response to lithium and may be associated with BPAD (Dimitrova et al., 2005). Glycogen kinase synthase 3 (GSK3) has numerous functions and some of its substrates play a role in neurone organisation and function (Plyte et al., 1992). Its activity is reduced by lithium and valproic acid, but not by carbamazepine (Klein and Melton, 1996; Stambolic et al., 1996).

A SNP in the promoter region of the *GSK3B* gene, which encodes glygogen synthase kinase 3 β, has been associated with later age of onset of BPAD (Benedetti et al., 2004a), better antidepressant effect of sleep deprivation (Benedetti et al., 2004b) and long-term response to lithium in BPAD (Benedetti et al., 2005).

Antidepressants also affect NMDA receptor function (Mjellem et al., 1993), and the expression of multiple genes encoding NMDA receptor subunits (Boyer et al., 1998). An allele of a variable microsatellite locus in the promoter of the *NMDA receptor 2A subunit* (*GRIN2A*) gene on 16p13 reduces transcription of *GRIN2A* (Itokawa et al., 2003b), and longer alleles at this locus are preferentially transmitted to BPAD patients (Itokawa et al., 2003a).

As for affective disorders, the identification of genes for susceptibility to SCZ by linkage analysis and association studies has also been successful. These genes and relevant related polymorphisms are described in the following section.

1.2.4 Genetic basis of schizophrenia

1.2.4.1 Positional candidate genes

Several chromosomal regions have been linked to SCZ: 1q21-22, 1p, 3q13.3, 5q, 6p22-24, 8p21-22, 10q22, 11q, 13q14-q21, 13q32, 15q14, 15q15, 18p, 22q11 (O'Donovan et al., 2003b; Lewis et al., 2003).

1.2.4.1.1 22q11 and COMT

Deletion and duplications of 1.5 to 3.0 Mb at 22q11 cause the chromosome 22q11 deletion syndrome, a variable phenotype including congenital heart defects, hypocalcemia, neurodevelopmental delays, facial dysmorphologies, cognitive deficits, and abnormal behaviours (Botto et al., 2003). The disease is also associated with a high rate of psychosis, including mood swings and SCZ (Pulver et al., 1994; Lachman et al., 1996a; Carlson et al., 1997; Bassett et al., 1998; Murphy et al., 1999; Murphy, 2002). Moreover, this region has been linked to BPAD (Kelsoe et al., 2001; Badner and Gershon, 2002) and SCZ (Karayiorgou et al., 1995; Blouin et al., 1998; Shaw et al., 1998; Badner and Gershon, 2002; DeLisi et al., 2002; Lewis et al., 2003; Williams et al., 2003b) and deletions are more frequent in schizophrenics than in unaffected people (Karayiorgou et al., 1995; Arinami et al., 2001), in particular with patients with early onset (Usiskin et al., 1999). Patients with chromosome 22q11 deletion syndrome and schizophrenics also share brain structural abnormalities (Chow et al., 1999; Chow et al., 2002).

The *COMT* gene located in this region is also a functional candidate due its involvement in the function and metabolism of monoamine neurotransmitters, in particular dopamine (Mannisto and Kaakkola, 1999; Harrison and Weinberger, 2005). A coding SNP in *COMT* increases the activity of the enzyme (Lotta et al., 1995; Lachman et al., 1996b) by a two-fold factor (Chen et al., 2004a; Shield et al., 2004). This polymorphism has been associated with SCZ by several independent studies and seems to have a small effect (Li et al., 1996; Kunugi et al., 1997; Li et al.,

1998c; Li et al., 1998d; Egan et al., 2001; Glatt et al., 2003a; Chen et al., 2004c). Other polymorphims have been identified in *COMT*. An association with SCZ was found in an isolated population for two SNPs, located in the upstream and downstream untranslated regions and both with small effect, and for a high frequency haplotype containing these SNPs and the aforementioned coding SNP (Shifman et al., 2002). Although these promising results remain to be replicated (Harrison and Weinberger, 2005), quantitative measures of allele-specific expression using mRNA from human brain suggest that the three-SNP haplotype is also associated with a 20% reduction in expression of *COMT* (Bray et al., 2003b). COMT also seems to be implicated in frontal lobe activity, characteristic of SCZ (Egan et al., 2001). Interestingly, a five-SNP haplotype spanning a 80 Kb region suggests that two genes in the 22q11 region, *proline dehydrogenase (oxidase) 1* (*PRODH*) and *zinc finger, DHHC domain containing 8* (*ZDHHC8*), may also be involved in SCZ (Liu et al., 2002b; Liu et al., 2002a).

1.2.4.1.2 22q11, PRODH, and ZDHHC8

PRODH is a mitochondrial enzyme which catalyses the first step in the catabolism of proline, which may directly modulate glutamatergic activity in the brain (Phang et al., 2001). Prodh mutant mice have reduced levels of glutamate and GABA, and also show abnormal sensorimotor gating, which is an endophenotype of SCZ (Gogos et al., 1999). Evidence suggesting the involvement of PRODH in SCZ comes from associations studies (Li et al., 2004a; Liu et al., 2002b), although three studies failed to replicate association of the *PRODH* gene with schizophrenia (Fan et al., 2003b; Williams et al., 2003a; Ohtsuki et al., 2004). A potential mechanism in the susceptibility to SCZ may involve the action of proline in modulation of glutamatergic transmission (Fremeau, Jr. et al., 1992; Renick et al., 1999) and apoptosis (Cohen and Nadler, 1997; Nadler et al., 1992). Mutations in PRODH may cause disturbance of neuronal function and hence predispose to SCZ (Karayiorgou and Gogos, 2004).

One of the SNPs on 22q11 originally associated with SCZ regulates the retention of intron 4 of *ZDHHC8* (Liu et al., 2002a; Liu et al., 2002b). This results in higher levels of unspliced variant thought to encode a truncated protein, thereby

modulating the expression of the fully functional transcript (Mukai et al., 2004). The *ZDHHC8* gene is thought to encode a transmembrane palmitoyltransferase, which seems to be involved in exploratory activity in mice (Karayiorgou and Gogos, 2004; Mukai et al., 2004). Palmitoylation is a reversible posttranslational addition of the saturated fatty acid lipid palmitate to proteins, which facilitate interactions between proteins and lipid bilayers (el Husseini and Bredt, 2002). Substrates of palmitoyltransferase include neuronal proteins that are important for neuronal development, synaptic scaffolding proteins, and neurotransmitter receptors, implying that synapse transmission may involve palmitoylation (el Husseini and Bredt, 2002). This is supported by findings that zinc finger, DHHC domain containing 3 (ZDHHC3) palmitoylates gamma 2 subunit of $GABA_A$ (Keller et al., 2004). The role of ZDHHC8 is not yet known, but the human brain expresses ZDHHC8 widely. The cortex and hippocampus, which are important in the pathophysiology of SCZ, show higher expression of ZDHHC8 than other regions (Mukai et al., 2004). Furthermore, ZDHHC8 deficiency correlated with decreased locomotor activity and greater fear of novel environment, in female, but not in male mice, which may involve glutamatergic transmission (Mukai et al., 2004). In humans, the aforementioned SNP shows distortion of transmission between sexes in schizophrenics (Mukai et al., 2004). Sexually dimorphic effects of ZDHHC8 might be related to differences observed between males and females in onset, incidence and severity of SCZ symptoms (Chen et al., 2004b; Mukai et al., 2004). However, follow up studies have failed to confirm this association (Saito et al., 2005).

In addition, the SCZ linked 22q12 region, contains several apoliprotein L (APOL) genes (Duchateau et al., 2001; Page et al., 2001). A 2.6-fold upregulation of *apolipoprotein L1* (*APOL1*), *apolipoprotein L2* (*APOL2*) and *apolipoprotein L4* (*APOL4*) has been observed in the prefrontal cortex of SCZ patients compared to that of controls (Mimmack et al., 2002). APOL proteins play a crucial role in cholesterol transport. The cholesterol content of membranes is important in cellular processes such as modulation of gene transcription and signal transduction during neurodevelopment and in the adult brain (Nimpf and Schneider, 2000).

1.2.4.1.3 13q34 and DAOA

Report of linkage to a gene desert on 13q with SCZ (Badner and Gershon, 2002) led to the annotation of another candidate oxidase gene, the gene encoding D-amino acid oxidase (DAO) activator (DAOA, also G72), and the *G30* gene, antisense of *DAOA* (Chumakov et al., 2002). Haplotypes at this locus have been associated with SCZ (Chumakov et al., 2002; Schumacher et al., 2004; Zou et al., 2005). In addition, four SNPs in the *DAO*, of which the product is activated by DAOA, have also been associated with SCZ (Chumakov et al., 2002). SNPs in the *DAOA* gene also seem to correlate with abnormal physiology and cognition in the prefrontal cortex and hippocampus of schizophrenics (Harrison and Weinberger, 2005). There is also evidence for a combined action of SNPs and haplotypes in these two genes (Chumakov et al., 2002; Schumacher et al., 2004). These findings support the role in SCZ of glutamatergic transmission via NMDA receptors because the substrate of DAO, D-serine, is an activator of NMDA glutamate receptors (Heresco-Levy et al., 1996; Mothet et al., 2000).

1.2.4.1.4 6p21 and NOTCH4

Several linkage analyses have suggested that 6p is linked to SCZ (Wang et al., 1995). Two polymorphisms in *NOTCH4*, which lies in this region, may be candidate susceptibility variants: a SNP in the promoter and a microsatellite in the first exon (Wei and Hemmings, 2000). Wei and Hemmings (2000) performed a transmission disequilibrium test (TDT) in the major histocompatibility complex (MHC) region at 6p21.3 in the attempt to identify a susceptibility gene in the region. The 13 markers tested spanned 1.8 Mb. Significant association was obtained for a locus corresponding to the *NOTCH4* gene. Further analysis of five markers identified two haplotypes strongly associated with SCZ, involving two SNPs in the promoter and a microsatellite in the first exon. This microsatellite encodes leucine in the signal peptide domain of the protein and an allele at this locus was transmitted in excess to affected offspring (Wei and Hemmings, 2000). Except for one study reporting that NOTCH4 is associated with SCZ (Glatt et al., 2005), subsequent studies failed to replicate these findings in several populations, including the population used in the

original positive report (Sklar et al., 2001; McGinnis et al., 2001; Wassink et al., 2003; Anttila et al., 2003; Tochigi et al., 2004).

However, while the microsatellite in exon 1 of the *NOTCH4* gene did not seem to be associated with SCZ, this polymorphism has been reported to be associated with endophenotypes for SCZ such as measures of cognitive performance in the frontal lobe and volumes of the frontal lobe in the brain (Wassink et al., 2003). Similarly, the SNP in the promoter of the *NOTCH4* gene failed to show association to SCZ, but showed strong association with earlier age of onset in male SCZ patients (Anttila et al., 2003). In addition, a microsatellite in the promoter of this gene, upstream of the two SNPs mentioned above and tested in the original report, has been associated with SCZ in an African sample, but not in samples of other ethnic backgrounds (Skol et al., 2003). These polymorphisms may also be associated with early-onset SCZ and SCZ with negative symptoms (Takahashi et al., 2003).

Recently, association studies of the two SNPs in the promoter of *NOTCH4* that have been associated with earlier age of onset, and SCZ with many negative symptoms, showed significant association of one haplotype with SCZ in African American but not European American populations. Furthermore the two groups differed in frequencies of alleles and haplotypes (Luo et al., 2004). These studies indicate that variants in NOTCH4 tested may be in LD with functional variants (Tochigi et al., 2004). They may also be population specific and have small effects (Takahashi et al., 2003; Luo et al., 2004; Zhang et al., 2004). Also, analysing homogenous samples via sample stratification based on the age of onset of the disease for example, and endophenotypes, rather than the disease status itself appears to increase the ability to detect association. Finally, polymorphisms in *NOTCH4* and *COMT* have been reported to be associated with response to antipsychotics and the combination of genotype at a locus in each gene correlate with a higher than 10 fold difference in the risk of being non-responsive (Anttila et al., 2004).

NOTCH4 is a member of the NOTCH family of transmembrane proteins, which mediate cell-cell interactions, controlling cell fate and thereby participating in a variety of developmental processes (Hansson et al., 2004; Schweisguth, 2004). Notch signalling is implicated in cell fate and is expressed in a wide variety of tissues, including the developing nervous system where it controls dendrite growth

and formation of neuronal contacts (Sestan et al., 1999). NOTCH4 is required during embryogenesis of the fruit fly for the correct segregation of epidermal cells from the neuronal cell precursors (Fortini and Artavanis-Tsakonas, 1993). In mice, NOTCH4 has been detected in the developing brain (Uyttendaele et al., 1996) and is involved in the structure and patterning of developing vasculature of embryos, including the brain (Uyttendaele et al., 2001). The role of NOTCH4 therefore remains undetermined, in particular in SCZ.

1.2.4.1.5 6p24 and DTNBP1

The short arm of chromosome 6 contains another positional candidate gene: dysbindin (dystrobrevin binding protein, DTNBP1). The *DTNBP1* gene is located on the linked region 6p24-22 (Straub et al., 1995) and has been associated with SCZ (Straub et al., 2002). DTNBP1 is present in the synapse termini of many neurones across the brain (Straub et al., 2002) and may mediate trafficking of NMDA, nicotinic and $GABA_A$ receptors, and of signal transduction proteins (Husi et al., 2000; Inoue and Okabe, 2003). A rare risk haplotype has been identified in an Irish sample which spans intron 2 to 5 of the *DTNBP1* gene (van den Oord et al., 2003). Another study of Irish sample did not support these findings (Morris et al., 2003). Subsequent studies of samples from different populations reported association of a common haplotype in introns 2-5 in samples from Germany, Hungary and Israel (Schwab et al., 2003), as well as association for different SNPs in a sample from England (Williams et al., 2004a), for a common haplotype in a Chinese sample (Tang et al., 2003), and for yet another haplotype in a Swedish sample (Van Den Bogaert., 2003). The existence of different risk haplotypes in different samples, including samples of similar geography and ancestry, suggest that several mutations with small effect have occurred in the *DTNBP1* gene leading to allelic heterogeneity (Harrison and Weinberger, 2005). One of these SNPs has been associated with difference in the expression of the *DTNBP1* gene (Bray et al., 2003a). The role of functional non-coding polymorphisms, including potential variants altering splicing and translation of the gene, is further supported by finding of lower levels of DTNBP1 mRNA and proteins in schizophrenics than in unaffected individuals

(Weickert et al., 2004), and the absence of coding variants in the *DTNBP1* gene (Harrison and Weinberger, 2005).

1.2.4.1.6 8p12 and NRG1

A recent study of a sample from Iceland has reported linkage and association of SCZ with the 1.4 Mb long *Neuregulin* (*NRG1*) gene on 8p12-21 (Stefansson et al., 2002), previously linked to SCZ (Badner and Gershon, 2002; O'Donovan et al., 2003b; Lewis et al., 2003). Although no single SNP was associated, a rare and low risk seven-SNP haplotype was associated which spanned 290 Kb of the upstream region up to the first exon of *NRG1* (Stefansson et al., 2002). In the same study, support for a role of the NRG1 in the aetiology of SCZ was obtained via $NRG1^{+/-}$ transgenic mice with halved NRG1 levels which showed abnormal behaviour akin to SCZ in mouse, such as hyperactivity triggered by exposure to novel environment (Stefansson et al., 2002). Association of NRG1 with SCZ has subsequently been reported in samples from several geographic origins, such as Scotland for the original haplotype (Stefansson et al., 2003), England and Ireland (Williams et al., 2003c; Corvin et al., 2004), and China (Yang et al., 2003; Li et al., 2004b; Zhao et al., 2004) for different haplotypes at the original SNPs (Williams et al., 2003c; Yang et al., 2003), and haplotypes at other SNPs (Tang et al., 2004; Corvin et al., 2004; Li et al., 2004b; Zhao et al., 2004). These findings strongly suggest that the *NRG1* gene manifests allelic heterogeneity, with variants clustered in two regions; one in the 5' regulatory region and one further downstream (Harrison and Weinberger, 2005). NRG1 has various roles mediated via several signal transduction pathways during development of the brain and periphery, such as neuronal migration and cellular differentiation, and neurotransmission and synaptic plasticity (Falls, 2003). These functions are mediated by the 15 NRG1 isoforms which are expressed in various areas of the brain and localised in several places in neurones (Law et al., 2004). NRG1 is present in glutamatergic synaptic vesicles and interacts with 'v-erb-b erythroblastic leukemia viral oncogene homolog' (ErbB) receptors to modulate NMDA receptor expression (Stefansson et al., 2002; Corfas et al., 2004). Although weak imbalance of NRG1 isoforms was reported in schizophrenics (Hashimoto et al.,

2004), this finding has not yet been replicated and no functional variants have been identified (Harrison and Weinberger, 2005).

In addition, haplotypes associated with modulation by nicotine of attentional and sensory processing and possibly SCZ have also been identified in the promoter the *α7 nicotinic receptor* (*CHRNA7*) gene (Leonard et al., 2002), in the 15q13-14 region linked to SCZ (Freedman et al., 2001). Several of these haplotype SNPs were functional in a gene reporter assay (Leonard et al., 2002). Although these findings are not yet replicated, they are promising because CHRNA7 is recruited to the synapse by NRG1 (Yang et al., 1998) and modulates dopamine and glutamate signalling pre- and post-synaptically (George et al., 2000; Fabian-Fine et al., 2001; Frazier et al., 2003; Paterson and Nordberg, 2000).

1.2.4.1.7 1q21, KCNN3 and RGS4

Several studies have reported linkage of 1q21-22 to SCZ (Brzustowicz et al., 2000; Gurling et al., 2001). Also, anticipation has been suggested by comparison of age of onset between affected probands and parents, which could be due to a gene with trinucleotide repeat expansions or other repeats affecting the expression of that gene (Stober et al., 1995; O'Donovan et al., 1996; Bassett and Husted, 1997; McInnis et al., 1999b; Vincent et al., 2000b). Genes with unstable trinucleotide repeats are therefore candidate genes. At 1q21.3, the '*small-conductance calcium-activated channel, subfamily N, member 3*' (*KCNN3*) gene is a candidate gene not only because it is involved in neuronal signalling but also because it contains two stretches of trinucleotide repeats coding for polyglutamine, one being highly polymorphic with allele size ranging from 4 to 28 repeats (Chandy et al., 1998; Gargus et al., 1998; Laurent et al., 1998; Antonarakis et al., 1999; Austin et al., 1999; Bonnet-Brilhault et al., 1999; Laurent et al., 2003). Expanded CAG repeat in *KCNN3* have been associated with ataxia (Figueroa et al., 2001) and anorexia nervosa (Koronyo-Hamaoui et al., 2004). Calcium-activated potassium channels are expressed in the brain and are involved in memory and learning (Stackman et al., 2002). They control calcium-mediated events via the regulation of neurone membrane by modulating the pacemaker duty cycle, in particular by creating slow after-hyperpolarisations, and spike frequency (Faber and Sah, 2003; Sah and Davies, 2000). Small-conductance calcium-activated potassium channels open upon increase

in concentration of cytosolic calcium. These channels are formed from homo- or hetero-dimers of subunits KCNNC1-4 (Faber and Sah, 2003).

KCNN3 is predominantly expressed in areas of the brain containing dopaminergic neurones (Dror et al., 1999; Rimini et al., 2000), involved in reward, working memory and voluntary movement (Goldman-Rakic, 1999; Kitai et al., 1999; Spanagel and Weiss, 1999). In dopamine neurones, blockage of these channels inhibits post-spike after-hyperpolarisation, changing the pacemaker-like discharge into multiple bursts, which are associated with increased release of dopamine (Ping and Shepard, 1999; Morikawa et al., 2000) and serotonin (Dawson and Routledge, 1995). Long alleles at the coding microsatellite locus have been reported to be associated with BPAD (Chandy et al., 1998) and also with SCZ, in particular negative symptoms (Bowen et al., 1998; Cardno et al., 1999; Dror et al., 1999; McInnis et al., 1999a; Ritsner et al., 2002; Ritsner et al., 2003). Many studies however did not replicate these findings (Laurent et al., 1998; Li et al., 1998b; Antonarakis et al., 1999; Bonnet-Brilhault et al., 1999; Hawi et al., 1999; Rohrmeier et al., 1999; Wittekindt et al., 1999; Chowdari et al., 2000; Stober et al., 2000; Saleem et al., 2000; Ujike et al., 2001; Laurent et al., 2003), or revealed association of short alleles (Stober et al., 1998b; Joober et al., 1999). However, a patient with SCZ had a 4-bp deletion in the CAG repeat region, creating a frameshift thought to result in a truncated protein (Bowen et al., 2001), and which erroneously localises to the nucleus, conferring dominant suppression of channel conductance and firing pattern involving KCNN2 (Miller et al., 2001). Longer alleles at the coding microsatellite might predispose to disease by stabilising the channel because they result in longer polyglutamine tracts, and thereby enhance the channel function (Gargus et al., 1998; Vincent et al., 2000b), in keeping with important discharge regularisation of dopaminergic neurones in animal models of SCZ (Svensson, 2000), but in contrast to the effects of apamin (Ping and Shepard, 1999) and the effect of the 4-bp deletion (Miller et al., 2001). These conflicting results and interpretations, as well as negative reports may reflect the fact that most studies tested association of the length of the allele with the disease, while some studies, including some that generated negative results for SCZ, detected association of negative and paranoid symptoms and early age of onset of SCZ, with the difference between allele sizes,

rather than the allele size itself (Cardno et al., 1999; Saleem et al., 2000; Ritsner et al., 2002). These findings are consistent with the heterodimeric association of the calcium-activated potassium channels, disruption of which might be the mechanism for the role of KCNN3 variation in SCZ (Saleem et al., 2000).

Similarly, haplotypes spanning the upstream region up to intron 1 of the *regulator of G-protein signalling 4* (*RGS4*) gene have been associated in several populations by several studies (Chowdari et al., 2002; Williams et al., 2004b; Morris et al., 2004). Although no functional variants have been identified (Chowdari et al., 2002), RGS4 mRNA levels are lower in SCZ patients than in controls (Mirnics et al., 2001), and *RGS4* is also an attractive functional gene since it is regulated by the dopaminergic pathway (Chowdari et al., 2002), inhibits G-protein mediated signalling via dopamine and glutamate receptors (De Vries et al., 2000; Ross and Wilkie, 2000), and is implicated in neuronal differentiation (Grillet et al., 2003).

1.2.4.1.8 1q42 and DISC1/DISC2

Linkage between a t(1;11)(q43;q21) reciprocal translocation and major psychiatric illness, including recurrent major depression, BPAD and schizophrenia was found in a Scottish family (St Clair et al., 1990; Muir et al., 1995). Evidence for linkage to 1q42 independent of the t(1;11) was reported in studies of the Finnish population (Hovatta et al., 1999; Ekelund et al., 2001; Hennah et al., 2003; Ekelund et al., 2004) and the Taiwanese populations (Hwu et al., 2003). A detailed haplotype map of the DISC locus was constructed based on previously and newly described SNPs, and these were tested for association with SCZ, schizoaffective disorder and BPAD (Thomson et al., 2005). Evidence for association was found for all three illnesses but not for a risk haplotype in common with those reported earlier in the Finnish (Hennah et al., 2003), Taiwanese (Hwu et al., 2003) or North American population (Hodgkinson et al., 2004). Further evidence for the existence of a candidate locus in this region comes from report of linkage and association of a polymorphic marker in the 1q42 region with spatial working memory, an endophenotype of SCZ, obtained by QTL analysis in schizophrenic individuals and their unaffected co-twins via QTL analysis (Gasperoni et al., 2003). There is as yet no convincing evidence for functional SNP association in *DISC1* despite evidence for association in several independent studies. It remains to be seen whether

functional SNPs in linkage disequilibrium with the associated markers will be found and whether there will be a common ancient variant or multiple risk variants relating to the multifunctional role of DISC1. In *DRD2*, a synonymous SNP alters the stability of the encoded mRNA and the level of expression of the DRD2 protein, dependant on the allele at another synonymous SNP locus in DRD2 (Duan et al., 2003). Also, non-coding SNPs in the *ZDHHC8* (Mukai et al., 2004) and *DTNBP1* (van den Oord et al., 2003; Owen et al., 2004) were reported to be associated with SCZ. In a similar fashion, the functional variant(s) in *TRAX/DISC1* may be non-coding, in regulatory regions or splice sites (Thomson et al., 2005). On the other hand, DISC1 is involved in various functions via interactions of different domains to several protein families suggesting that DISC1 function could be influenced by allelic heterogeneity (Harrison and Weinberger, 2005). It would therefore be valuable to integrate knowledge on functional domains in proteins and functional conservation across species (Thomson et al., 2005).

1.2.4.2 Functional candidate genes

As for affective disorders, drugs with a beneficial effect on SCZ symptoms, such as clozapine, affect the function of neurotransmitters or their receptors. Genes implicated in the dopaminergic and serotonergic systems, and the γ–butyric acid (GABA) pathway are therefore potential candidate genes (Harrison and Weinberger, 2005). In addition, histological observations of abnormal localisation and clustering of neurones in the brain of schizophrenics suggest the existence of mutations in genes implicated in neuronal migration, survival and connectivity (Roberts, 1990; Harrison and Weinberger, 2005; Harrison, 1997).

1.2.4.2.1 Serotonergic pathway

The role of serotonin in SCZ is supported by the observation that serotonin receptors and sites for reuptake are altered in the limbic system in schizophrenics (Joyce et al., 1993). Furthermore, the *HTR2A* gene at 13q14 may be associated with affective symptoms (Fanous et al., 2004) and has been associated with response to clozapine (Arranz et al., 1998). Also, the VNTR in intron 2 of the *serotonin transporter* (*SCL6A4*) gene described above has been associated with SCZ (Tsai et al., 2002; Kaiser et al., 2001; Liu et al., 1999), while the 44 bp indel in the promoter

has also been associated with schizoaffective and schizoparanoid subtypes (Kaiser et al., 2001) and endophenotypes of SCZ (Malhotra et al., 1998), but not susceptibility to SCZ (Rao et al., 1998; Tsai et al., 2002). In addition, genotypes at the *SLC6A4* promoter variant and at loci in *HTR2A* and *5HT2C* have been used to predict response to clozapine in schizophrenics (Arranz et al., 2000; Malhotra et al., 2004; Johnson et al., 2005).

1.2.4.2.2 Dopaminergic pathway

The efficiency of antipsychotic drugs is correlated with affinity for dopamine D2 receptors, suggesting brain dopamine hyperactivity in SCZ (Malhotra et al., 2004). Typical antipsychotic drugs which block DRD2 are effective in treating positive symptoms, but less so for negative symptoms (Kane and Freeman, 1994). DRD2 is located on post-synaptic dopaminergic neurons involved in reward-mediating pathways, locomotion, hormone production and drug abuse, and is the target of drug to treat psychoses, such as haloperidol for SCZ (Seeman, 1987; Neville et al., 2004). Several polymorphisms in *DRD2* have been associated with SCZ in samples with different ethnic origin: i) a functional single nucleotide insertion/deletion in the promoter in some (Arinami et al., 1994; Ohara et al., 1998; Inada et al., 1999; Jonsson et al., 1999) but not in all studies (Li et al., 1998a; Glatt et al., 2004), ii) a SNP in a coding exon (Shaikh et al., 1994; Arinami et al., 1996; Arinami et al., 1997; Glatt et al., 2003b), also associated with delusions in major psychoses (Serretti et al., 2000a), iii) two TaqI variants in intron 2 (Dubertret et al., 2001; Golimbet et al., 2003) and iv) a microsatellite in intron 2 (Golimbet et al., 2003), which is also associated with Parkinson's disease and with levodopa treatment-induced dyskinesia in patients (Oliveri et al., 1999).

Many drugs used to treat SCZ are dopamine receptor antagonists and can have secondary effects in schizophrenics such as rigidity, bradykinesia, dyskinesia (Malhotra et al., 2004). The more recent drug clozapine is used to treat positive and negative symptoms (Leucht et al., 2003). Clozapine has a 10-fold greater affinity for DRD4 than for DRD2 suggesting an important role for DRD4 in SCZ (Van Tol et al., 1991). A SNP located at –521 with respect to the initiation codon reduces expression of the *DRD4* gene by 40% in transient expression assay, with the high-expression

allele being associated with SCZ (Okuyama et al., 1999), and novelty seeking (Okuyama et al., 2000; Ronai et al., 2001). Similarly, a 120 bp tandem repeat polymorphism located 1.2 Kb upstream of the initiation codon has been associated with SCZ (Xing et al., 2003), ADHD (McCracken et al., 2000; Mill et al., 2003), methamphetamine abuse (Li et al., 1997b), and novelty seeking in bipolar and alcoholic families (Rogers et al., 2004). Haplotypes at the VNTR and SNPs closer to the gene, including the functional SNP at –521, have also been associated with SCZ (Xing et al., 2003). The 120 bp duplicated sequence contains binding sites for several transcription factors and polymorphisms within these sites may affect regulation of transcription of *DRD4*, as it has been observed in gene reporter assays in human cells where the long allele showed lower expression (D'Souza et al., 2004). Mice lacking *DRD4* are less active in open fields tests, but are more active on rotarod and display increased locomotor sensitivity to ethanol, cocaine, methamphetamine (Rubinstein et al., 1997).

Polymorphisms in other genes of the dopamine pathway have been associated with SCZ. A microsatellite, named DRD5 and located 18.5 Kb upstream to the *DRD5* gene has been associated with SCZ (Williams et al., 1997; Muir et al., 2001) and ADHD in some (Daly et al., 1999; Tahir et al., 2000; Maher et al., 2002; Kustanovich et al., 2004; Lowe et al., 2004), but not all studies (Barr et al., 2000; Payton et al., 2001) but not with BPAD nor UP (Muir et al., 2001). This marker is also thought to be in LD with one or more causal variants and variation in LD between these variants may confound results of these association studies (Daly et al., 1999; Muir et al., 2001; Mill et al., 2004). Several non-synonymous mutation have been identified in transmembrane domains and intra-and extra-cellular loop of DRD5, some of which resulted in small to 10-fold increase in dopamine affinity (Cravchik and Gejman, 1999). A microsatellite, D4S615, located 131 Kb downstream of the *DRD5* gene has been associated with SCZ (Muir et al., 2001). A microsatellite in a trans-activation domain in the promoter of *DRD5* has been associated with dystonia, which is often co-morbid with SCZ, by some studies (Placzek et al., 2001; Misbahuddin et al., 2002). These findings were not replicated by other studies (Sibbing et al., 2003). Because of conflicting results obtained by

linkage and association studies of *DRD5*, it has been suggested that the markers tested may be in LD with variants in other genes (Muir et al., 2001).

Polymorphisms in the *DRD3* gene have been associated with SCZ (Crocq et al., 1992; Spurlock et al., 1998; Lohmueller et al., 2003), but not by all studies (Anney et al., 2002). DRD3 deficient mice showed hyperactivity and increased rearing behaviour (Accili et al., 1996). A non-synonymous mutation in the *DRD3* gene has also been associated with eye movement disturbances, which are used as a marker for SCZ (Holzman, 2000) and could be used as a marker itself (Rybakowski et al., 2001). A rare allele at a microsatellite locus in the first intron of the *tyrosine hydroxylase* (*TH*) gene, implicated in the synthesis of dopamine, has been associated with SCZ in some studies (Meloni et al., 1995) but not by others (Jonsson et al., 1998; Burgert et al., 1998). A 40bp VNTR polymorphism in the 3' untranslated region of the *dopamine transporter* (*SLC6A3*) gene may be associated with negative symptoms, in keeping with the possible role of abnormal prefrontal dopamine function in negative symptoms (Goldman-Rakic and Selemon, 1997), the modulation of the activity of dopamine at all types of dopamine receptors by SLC6A3 (Fanous et al., 2004), and the amelioration of negative symptoms by clozapine treatment, possibly via activation of dopamine release in the prefrontal cortex (Youngren et al., 1999).

1.2.4.2.3 Neurotrophins

Neurotrophins may play an important role in SCZ as they are crucial for neurodevelopment (Labelle and Leclerc, 2000; Murer et al., 2001). For example, brain derived neurotrophic factor (BDNF) has been associated with SCZ (Muglia et al., 2003). BDNF promotes survival of neurons in the hippocampus and cerebral cortex, which both express *BDNF* in adults brains (Murer et al., 2001), axonal outgrowth (Labelle and Leclerc, 2000) and synaptic plasticity (Ying et al., 2002). BDNF mediates long-term neuronal adaptations by governing the response to dopamine of target neurones, by modulating the expression of *DRD3* during development and adult life, for example (Guillin et al., 2001). Together with BDNF, neurotrophin-3 (NT3) activate a phosphatidylinositol (PI) 3-kinase (PIK3C3), thereby generating PI(3)P and activating AKT1 (Yoshimura et al., 2005). GSK3B is

inactivated by 'v-akt murine thymona viral oncogene homolog 1' (AKT1) in response to neurotrophins (Yoshimura et al., 2005). GSK3B appears to regulate the activity of microtubule-associated proteins. Its inactivation may promote axon outgrowth (Yoshimura et al., 2005). Furthermore, the decrease in levels of phosphorylation of GSK3B and the association of a haplotype in *AKT1* with lower AKT1 proteins levels observed in the prefrontal cortex and hippocampus of schizophrenics suggest that AKT1 may be involved in SCZ (Emamian et al., 2004). AKT1 is a multifunctional kinase that mediates multiple responses via several signalling pathways, and numerous protein-protein interactions (Hallmayer, 2004). Implication of AKT1 in the aetiology of SCZ is further supported by greater sensitivity to disruption of sensorimotor gating by amphetamine in *Akt1* knockout mice, and by the amelioration of this phenotpe upon treatment with the drug haloperidol which is used to treat SCZ (Emamian et al., 2004).

1.2.4.2.4 Glutamatergic pathway

As shown by linkage analyses, genes in the glutamatergic system, such as *DAO*, are also potential candidates for susceptibility to SCZ, in particular glutamate receptors (Schiffer, 2002; Collier and Li, 2003). Haplotypes at two to five SNPs spanning intron 3-5 of the *glutamate receptor 3* (*GRM3*) gene have been associated with SCZ (Fujii et al., 2003; Egan et al., 2004). Intriguingly, a SNP in intron 2 is associated with SCZ, and with cognitive and brain activity defects which were observed not only in patients but in normal individuals too (Egan et al., 2004). Effects of GRM3 on the aetiology of SCZ may be mediated via its action in modulating dopamine, serotonin, and NMDA receptor (NMDAR) transmission (Cartmell and Schoepp, 2000; De Blasi et al., 2001; Spooren et al., 2003).

These findings show that although association of several genes with SCZ has been replicated, specific alleles have not, with risk alleles typically varying between studies (Harrison and Weinberger, 2005). This may be due to the fact that these studies rely on samples from populations with different ancestral genetic backgrounds and hence harbouring different haplotypes (Botstein and Risch, 2003; Schwab et al., 2003). Candidate genes for SCZ identified support the view that SCZ is a disorder of synaptic signalling, especially in dopaminergic and glutamatergic

signalling, including receptors such as DRD2, DRD4, GRM3, CHRNA7, DAO and DAOA (Harrison and Weinberger, 2005), as well as signal transduction, via RGS4 for example, and synapse plasticity and formation, involving NRG1, DNTBP1 and DISC1.

1.2.5 Genetic components common to affective disorders and schizophrenia

Although standardised criteria to define affective disorders and SCZ exist, symptoms for these three illnesses overlap widely (American Psychiatry Association, 1994). The existing classification of these illnesses remains uncertain due to the current lack of reliable biological and genetic markers, and the disorders may share genetic and nongenetic risk factors (Berrettini, 2000; Berrettini, 2004). This is supported by the existence of both illnesses in the same kindreds, together with intermediate phenotypes, such as schizoaffective disorder, and the transition from one phenotype to the other in some patients (Blackwood et al., 1996; Wildenauer et al., 1999). The co-localisation of regions linked to SCZ and BPAD, such as the DISC locus and 10p13-p12, 11q23, 13q32, 18p, 18q and 22q11-q13, further suggests that some genetic susceptibility factors are shared between the two diagnoses (Wildenauer et al., 1999; Berrettini, 2004).

Studies of functional candidate genes also indicate common pathways, such as the serotenergic pathway implicating genes harbouring polymorphisms associated with both psychoses, or symptoms of these illnesses, such as one of the serotonin receptors (Fanous et al., 2004; Malhotra et al., 2004), serotonin transporter (Rotondo et al., 2002), dopamine transporter (Greenwood and Kelsoe, 2003; Hawi et al., 2003; Fanous et al., 2004; Keikhaee et al., 2005), or a monoamine oxidase (Shifman et al., 2002; Shifman et al., 2004).

There is also evidence that genes implicated in glutamatergic transmission contribute to both diseases too, such as in the case of DAO and the activation of the NMDA-type glutamate receptor (Heresco-Levy et al., 1996). The *DAO* gene is located in a region linked to BPAD (Ewald et al., 2002; Shink et al., 2005) and has been associated with SCZ (Schumacher et al., 2004). The gene coding for the activator of DAO, *DAOA*, is located in 13q32-33, which shows linkage to SCZ and

BPAD (Badner and Gershon, 2002), and has been associated both with SCZ and BPAD (Chumakov et al., 2002; Hattori et al., 2003; Schumacher et al., 2004; Zou et al., 2005).

Similarly, neutrophins may contribute to both disorders. For example, a VNTR in the 3' untranslated region UTR of *BDNF* was reported to be associated with BPAD (Neves-Pereira et al., 2002), with SCZ (Muglia et al., 2003) and with negative and affective symptoms in SCZ (Fanous et al., 2004). This variant and a polymorphism in the upstream region of *BDNF*, reported to be associated with negative symptomes, only explains a small proportion of the variance in the SCZ symptoms reflecting the possibility that complex traits may involve several genes and interactions between these genes (Fanous et al., 2004). In addition, the AKT1/GSK3B signalling pathway, target of lithium and valproic acid, seems to be implicated in BPAD (Benedetti et al., 2005) while a polymorphism in the *AKT1* gene has been associated with SCZ (Emamian et al., 2004).

Apart from the polymorphism in the coding sequence of *COMT*, most polymorphisms found so far and associated with affective disorders and SCZ are non-coding. Although evidence is still lacking for the functional role of most of the non-coding variants, and these variants may therefore not contribute to the aetiology of psychosis, but rather be in linkage disequilibrium with coding variants, it is likely that many alter the expression of the genes (Harrison and Weinberger, 2005).

The sequence of the human genome has allowed fast progress in the identification of sequence variants. The next section introduces the Human Genome Project on which many findings mentioned above derived and on which the further elucidation of the molecular basis for genetic risk depends.

1.3 The Human Genome Project

The completion of the Human Genome Project conducted by the International Human Genome Sequencing Group is a landmark event and is having a profound impact on biomedical research (Collins et al., 2003). Knowledge of the human genome sequence permits the comprehensive determination of the structural and functional elements it encodes, the organisation of genetic networks and biochemical pathways, as well as the description of the genetic variation. This wealth of

information should in turn enable the elucidation of the interplay between genetic and environmental factors and how it translates into organismal phenotypes, such as disease and response to treatment (Davies, 2001; Collins et al., 2003). Four features of the Human Genome Project enabled important progresses in the understanding of complex disorders, such as psychiatric diseases: creation of physical maps, annotation of the genomic sequences, identification of numerous nucleotide polymorphisms and comparative genomics (Collins et al., 2003).

The publicly available sequence covering 90% of the human genome was first announced in February 2001 (Venter et al., 2001; International Human Genome Sequencing Consortium, 2001). It was freely available from GenBank in the form of unfinished and fragmentary BAC clone sequences (http://www.ncbi.nlm.nih.gov/Genbank/). Although the Human Genome Project was completed in April 2003, the second, 'finishing' phase is now completing sequence of the existing clones (National Human Genome Research Institute web site, http://www.nhgri.nih.gov). The assembly of the human genome sequence is regularly updated (http://www.ncbi.nlm.nih.gov/genome/human/release_notes.html) by incorporating these new sequences. The human genome shows variation in the distribution of CG content, recombination rate, CpG island, genes, and transposable elements. It also contains approximately 30,000 genes, that is only twice as many as in worm or fly (International Human Genome Sequencing Consortium, 2001). Most additional genes in human are however not entirely novel but instead derived from duplication of existing ones. Human genes also undergo more complex regulatory events, such as splicing, leading to a greater variety of proteins. Furthermore, human proteins seem to contain more diverse combinations of domains. (International Human Genome Sequencing Consortium, 2001; Levine and Tjian, 2003). Transposable elements comprise about half the human genome, but their activity appears to have declined dramatically (International Human Genome Sequencing Consortium, 2001). Also, most mutations appear in males, as shown by a mutation rate in male meiosis being twice that of female meiosis (International Human Genome Sequencing Consortium, 2001). Finally, at least one crossover occurs per chromosome arm in each meiosis and recombination rates tend to be higher in distal regions of

chromosomes and on short arms (International Human Genome Sequencing Consortium, 2001).

The assembly of the human genome generated by the National Center for Biotechnology Information (NCBI) and its annotation can be consulted online at the Entrez Genomic Biology page for Human (http://www.ncbi.nlm.nih.gov/genome/guide/human/). One can also access this sequence and the annotation generated by Ensembl, a joint project of the European Bioinformatics Institute and the Wellcome Trust (Hubbard et al., 2002), and that generated by the University of California at Santa Cruz (Kent et al., 2002). Finally, Celera Genomics also generated an assembled sequence, available from its online site under various restrictions (http://public.celera.com/index.cfm). With the completion of several genome sequencing projects and the development of new experimental approaches to study gene expression patterns and proteomics, the volume and complexity of biological data will continue to increase dramatically (Collins et al., 2003). These databases and related web sites have therefore become a tool important to many biologists. The successful completion of the Human Genome Project led to other genome projects for vertebrate model organisms, such as mouse (Waterston et al., 2002), rat (Rat Genome Sequencing Project Consortium, 2004), chimpanzee (Olson and Varki, 2003), dog (Kirkness et al., 2003) and chicken (Hillier et al., 2004). As sequences of genomes are released, they are incorporated into the aforementioned databases, enabling research progress within and between organisms (Clamp et al., 2003; Kent et al., 2002). Together with genome sequences of the fruit fly and nematode, this data directly helps to understand human diseases, as it appears that important genes involved in development and complex traits, such as behaviour, in these model organisms share conserved function with human genes. For example, transgenic mice with low expression level of the NMDA receptor showed abnormal behaviours akin to symptoms in schizophrenia, which were ameliorated upon treatment with haloperidol and clozapine which are used to treat schizophrenic patients (Mohn et al., 1999). While the Human Genome Project covered the entire genome, including the 99,9% of the bases shared by all humans, it also enabled the characterisation of the 0.1% remaining bases that vary between individuals.

1.4 Human genetic variation and Single Nucleotide Polymorphisms

Human genomic sequence variation includes single nucleotide, micro- and mini-satellite repeat, insertion and deletion polymorphisms (International Human Genome Sequencing Consortium, 2001). Single Nucleotide Polymorphisms (SNPs) are a powerful tool to uncover associations of specific loci with diseases (Risch and Merikangas, 1996; Bentley, 2000). They present two advantages over other types of markers. First, it is relatively easy to automate SNP typing assays which can be performed on a large-scale and remain cheap and robust (Gonen et al., 1999; Bartlett et al., 2001; Ranade et al., 2001; Xu et al., 2003). Second, SNPs are abundant throughout the genome and can be found on average every 100 to 300 bases (http://www.ncbi.nlm.nih.gov/SNP/). SNPs can be detected using targeted or non-targeted techniques. The latter approach consists of large-scale shotgun sequencing of libraries of total genomic DNA or individual chromosomes (Taillon-Miller et al., 1998). This method was used by a consortium of pharmaceutical companies, the Wellcome Trust and international genome centres: The SNP Consortium (TSC, http://snp.cshl.org/). The NCBI dbSNP database has been designed to act as a central repository for SNP data, including those from large projects such as TSC (Smigielski et al., 2000). The recent dbSNP build 122 (http://www.ncbi.nlm.nih.gov/SNP/) contains approximately 9.8 million SNPs, of which approximately 5 million are validated (Mooney, 2005). The increasing density of SNPs on the human genome facilitates genome-wide association studies and replication studies, as well as the understanding of variation in response to treatment observed between patients and the identification of population stratification (Roses, 2000; The International HapMap Consortium, 2003; Malhotra et al., 2004; Johnson et al., 2005). Estimates of the density of SNPs required for powerful genome-wide association studies has initially varied from 1 per 3 Kb to 1 per 100 Kb (Risch, 2000; Hirschhorn and Daly, 2005). Most functional human genetic variation appears not to be population-specific (Stephens et al., 2001; Hinds et al., 2005), although some examples that link population specific polymorphism and phenotypic differences exist (Xiao and Casero, Jr., 1996; Nakayama et al., 2002a; Nakayama et al., 2002b; Nakayama et al., 2002c; Nakayama et al., 2002d; Duggal et al., 2003).

As described above, the indirect approach for association studies is thought to be able to capture most human sequence variation with greater efficiency than the direct method for two reasons. First, common variants account for approximately 90% of inter-individual sequence variation (Kruglyak and Nickerson, 2001; Reich and Lander, 2001). Second, single historical events generated most of these variants, which are thus associated with neighbouring polymorphisms also present on the ancestral chromosomes (The International HapMap Consortium, 2003).

The aim of the HapMap Project is to characterise common patterns of sequence polymorphism in the human genome by determining the frequencies of alleles and genotypes at millions of loci across the genome and the association between them (The International HapMap Consortium, 2003). This data should allow any given disease to be tested indirectly for association with common variants spread across the entire genome, chromosomal regions identified by linkage analysis, or in and around functional candidate genes (The International HapMap Consortium, 2003). Identification of rare disease-causing variants will also benefit from this data as these polymorphisms will be linked to common variants enriched in cases compared to controls (The International HapMap Consortium, 2003), reminiscent of the identification of the *cystic fibrosis transmembrane conductance regulator* (*CFTR*) gene responsible for cystic fibrosis (Kerem et al., 1989). Most of the genetic variation observed in the world's human population is predicted to occur at approximately 10 million SNPs, namely one locus every 300 bp, where alleles have a frequency of at least 1%. The remaining 10% of the inter-individual genetic variation would be accounted for by rare variants (Kruglyak and Nickerson, 2001; Reich and Lander, 2001).

Empirical studies have shown that most of the human genome harbours a block-like structure with regions a high LD between nearby SNPs, separated by regions of low LD (Abecasis et al., 2001; Jeffreys et al., 2001; Dawson et al., 2002; Gabriel et al., 2002; Reich et al., 2002). The regions of high LD usually contain only a few haplotypes that jointly account for most of the observed variation (Rioux et al., 2000; Patil et al., 2001; Dawson et al., 2002). It follows that alleles at many loci within these regions of high LD can be deducted by genotyping a few linked SNPs only, termed 'tag' SNPs (Johnson et al., 2001; Goldstein et al., 2003). The practical

impact of these findings is that, accounting for the variable extent of LD in the human genome, between 200,000 and 1,000,000 tag SNPs are expected to capture most of the common genetic variation in human populations due to the 10 million common SNPs in the human genome. The small loss of information inherent to this deduction is outweighed by the large reduction in the amount of genotyping (The International HapMap Consortium, 2003). For common SNPs, which are thought to be older than rare SNPs, historical recombination and demographic events can account for the observed patterns of LD (Chakravarti, 1999). Some recombination events tend to be clustered in recombination hotspots (Chakravarti et al., 1984; Jeffreys et al., 2001). These findings are consistent with the observation that unrelated chromosomes and some populations shared haplotypes and patterns of LD (Gabriel et al., 2002). Elucidation of LD patterns in the human genome will contribute to the identification of polymorphisms involved in diseases, in particular functional variants in non-coding sequences, likely to be involved in gene expression (The International HapMap Consortium, 2003).

Gene expression can be altered by non-coding polymorphisms in the promoter and introns at the level of transcription (Greenwood and Kelsoe, 2003; Laws et al., 2002; Lemonde et al., 2003; Theuns et al., 2003). In introns, SNPs can alter transcript and isoform abundance via splicing or mRNA stability (Tokuhiro et al., 2003), and in the 3'untranslated region at the level of mRNA and translation (Mill et al., 2002; Miller and Madras, 2002). Variation in gene expression is presented in the following section.

1.5 Gene expression variation

1.5.1 Impact on morphology

Variation in gene expression may influence evolution more strongly than variation in protein isoforms (King and Wilson, 1975) and differences in gene expression account for a major part of the variation within and between species (Levine and Tjian, 2003). Polymorphisms in *cis*-acting regulatory regions have been implicated in morphological features in insects, such as the conversion of swimming limbs into feeding appendages in certain crustaceans, or the conversion of hindwings

in butterfly into rudimentary wings in Drosophila, due to mutations creating bindings sites for a repressor transcription factor Ultrabithorx (Ubx) (Weatherbee et al., 1999). Recently, novel regulatory elements in the fruit fly have been identified in existing regulatory regions and shown to affect the pigmentation of the abdomen and wings in males, and may therefore influence mating selection (Gompel et al., 2005). Similarly, analysis of sequences of the *homeobox C8* (*HOXC8*) gene, which controls axial morphology in vertebrates, from fish and mammals, led to the identification of functional nucleotide substitutions in non-coding regions, as well as species specific set of functional regulatory elements (Shashikant et al., 1998; Anand et al., 2003; Wang et al., 2004). In human, a mutation in the intron of the '*chromosome 7, open reading frame 2*' (*C7orf2* or *limb region 1*, *LMBR1*) gene, directly disrupts a *cis*-acting regulator of the '*sonic hedgehog homolog*' (*SHH*) gene located 1Mb away (Lettice et al., 2002). This variant may be implicated in limb malformation, namely Preaxial polydactyly (PPD), since mis-expression of the *Shh* gene in mouse is a common requirement for creating extra digit (Lettice et al., 2002).

1.5.2 Role in diseases

The major cause of disease known to date is mutation in coding regions of genes which affect protein quantity and/or quality (Cargill et al., 1999), as in Hungtington's disease for example, which is caused by trinucleotide repeat expansion leading to enlarged repetitive series of the encoded amino acid in the protein (The Huntington's Disease Collaborative Research Group, 1993). However, some of these mutations affect proteins involved in gene expression regulation, such as in the Rett syndrome for example (Caballero and Hendrich, 2005). Furthermore, polymorphisms in regulatory regions have been shown to increase risk of diseases (Risch and Merikangas, 1996; Collins et al., 1997; Wray et al., 2003; Knight, 2005).

At present, approximately 26,000 mis-sense and nonsense mutations are known, whereas only 545 regulatory mutations have been identified (Stenson et al., 2003). This difference is probably biased and may reflect the relative ease of identifying mutations in coding regions, mainly because they alter the structure of the protein, compared to the difficult characterisation of non-coding functional mutations (Knight, 2005). Non-coding mutations may also have smaller effects, as observed for

mutations with milder effects affecting complex traits (Toma et al., 2002), and, in isolation, may not be sufficient to cause disease (Botstein and Risch, 2003). Early examples of regulatory variants causing disease include substitutions in regulatory elements in the *β-globin* gene cluster implicated in β-thalassemia (Antonarakis et al., 1984), in the promoter of the *low density lipoprotein* gene involved in hypercholesterolemia (Koivisto et al., 1994), the *factor IX* gene in haemophilia (Picketts et al., 1994), the *chemokine receptor 5* gene in acquired immune deficiency syndrome (McDermott et al., 1998), a 2-bp insertion in the promoter of the *bilirubin UPD- glucuronosyltransferase* gene reducing the expression of the gene and causing the jaundice in the Gilbert syndrome (Monaghan et al., 1996), and splicing variation of CFTR variants in cystic fibrosis (Rave-Harel et al., 1997). Inherited variation in gene expression may thus play an important role in susceptibility to complex diseases (Lander, 1996; Peltonen and McKusick, 2001). Supporting this hypothesis, regulatory polymorphisms are increasingly recognised to contribute to complex disease traits, although the number of examples with both a clear-cut genetic association and defined molecular mechanism remain small (Knight, 2005). For example, the fragile X syndrome, characterised by mental retardation, is caused by alteration of the expression of the *fragile X mental retardation-1* (*FMR1*) gene due to expanded repeats in its promoter (Kenneson et al., 2001). Several other recent studies have reported non-coding polymorphisms, notably in the regulatory region of genes, associated with complex diseases, such as lupus erythomatosus (Werth et al., 2000), asthma (Moffatt and Cookson, 1997), diabetes (Horikawa et al., 2000b), inflammatory bowel disease (Horikawa et al., 2000a), autoimmune disease (Ueda et al., 2003), rheumatoid arthritis (Suzuki et al., 2003), myocardial infarction and stroke (Helgadottir et al., 2004), and schizophrenia (Wei and Hemmings, 2000; Stefansson et al., 2002; Straub et al., 2002; Bray et al., 2003b; Schwab et al., 2003).

1.5.3 Heritability

Variation in genome-wide gene expression has been associated with phenotypic variation in a number of different organisms: the budding yeast (Brem et al., 2002; Townsend et al., 2003; Fay et al., 2004), drosophila (Rifkin et al., 2003), teleost fish (Oleksiak et al., 2002), primate species (Enard et al., 2002) and human

(Pastinen et al., 2004; Knight, 2004; Yan et al., 2002; Cheung et al., 2003). Analysis of the intra- and inter-specific variation in primate gene expression patterns has shown that gene and protein expression have evolved extensively in the human brain, and that global differences in gene expression between the few individuals tested is as great as that observed between human and chimpanzee (Enard et al., 2002).

In human, recent studies have shown that gene expression variation is common for autosomal genes that do not show genomic imprinting (Rockman and Wray, 2002; Buckland, 2004). Out of 107 genes with variants in the 5' flanking region that were experimentally verified as functional in reporter gene assays in physiologically relevant cell lines, most polymorphisms were in the proximal promoter regions, but over a quarter were located more than 1 Kb upstream or 3' downstream to the transcriptional start site (Rockman and Wray, 2002). This study also revealed that 67% of the variants showed allelic differences in the rate of transcription of at least 2-fold, and 10% had differences of at least 10-fold. It was also estimated that 30% of genes have functional promoter SNPs (Rockman and Wray, 2002). In a different study of gene expression in human using two cell lines under basal conditions, 11% of 249 genes tested had promoter haplotypes showing differences in gene expression of at least 1.5, and more than half of these effects could be ascribed to variants with minor allele frequency less than 5%, suggesting that more functional variant would be found if more individuals, conditions and cell lines had been screened (Buckland, 2004). Recent *in vivo* studies of relative allelic expression in normal tissues or cell lines for hundreds of human genes suggest that 25% to 50% of genes and 5% to 25% of heterozygotes show evidence of unequally expressed alleles (Yan et al., 2002; Bray et al., 2003a; Lo et al., 2003; Pastinen et al., 2004). Finally, robust allele differences in gene expression appear to be of modest amplitude, typically 1.5 – 2.0 fold, in mouse (Cowles et al., 2002) and human (Yan et al., 2002; Bray et al., 2003a; Pastinen et al., 2004).

Some studies have also shown that gene expression variation is heritable in human (Yan et al., 2002; Cheung et al., 2003; Pastinen et al., 2004). For example, Cheung and co-workers reported that microarray analysis of 5000 randomly chosen cDNAs using human cell lines established from unrelated and related individuals allowed the identification of genes whose transcript levels varied most among normal individuals (Cheung et al., 2003). Valid observations on at least three out of four replicates in at least 30 out of 35 unrelated individuals were obtained for 813 genes.

Further analysis of the gene expression of five highly variable genes in unrelated individuals, offspring and monozygotic twins showed that the variance in gene expression among unrelated individuals was 3-11 times greater than that between monozygotic twins, and that the variance among sibling was 2-5 times greater than that between twins, suggesting the existence of genetic basis for gene expression variation (Cheung et al., 2003). Other studies provide evidence for the inheritance of some differences in gene expression in humans, and also show that difference in transcript abundance among non-imprinted genes support existence of regulatory polymorphisms (Yan et al., 2002; Pastinen et al., 2004).

1.5.4 Functional variants

Factors hindering the characterisation of the extent of *cis*-regulatory variation include: i) the current limited ability to recognise regulatory regions of most genes and to predict which of these variants might affect gene expression given the vast amount of polymorphisms in the human genome (International Human Genome Sequencing Consortium, 2001; Cowles et al., 2002), and ii) linkage disequilibrium between functional and non-functional variants (Knight, 2005). In addition, experimental screens for regulatory variation in human solely based on transcript levels are complicated by the possibility that such differences could also be caused by *trans*-acting and environmental factors (Cowles et al., 2002). Confounding *trans*-acting and environmental factors can be controlled for by studying F1 hybrids from crosses of two strains or populations so that both alleles are expressed in identical conditions. This was applied to yeast (Cohen et al., 2000; Brem et al., 2002), fruit fly (Wittkopp et al., 2004), fish (Shapiro et al., 2004) and mouse, where difference in expression of 1.5 fold or greater can be confidently detected (Cowles et al., 2002). In yeast, genetic linkage analysis of genomewide expression patterns in a cross of two yeast strains identified *cis*- and *trans*- modulators with a handful of *trans*-acting modulators each affecting the expression of a few genes to almost a hundred genes (Brem et al., 2002). Evidence that some closely linked genes are influenced by the same *cis*-regulator and have correlated expression profiles was also found in yeast (Cohen et al., 2000), in fruit fly (Spellman and Rubin, 2002) and in human (Caron et al., 2001).

Genome-wide linkage analysis of gene expression variation in human identified putative *cis*-regulators, and for some of them, confirmation of the association at a population and familial level was obtained by typing additional SNPs (Morley et al., 2004). On the other hand, *trans*-acting regulatory variation appears to be the source of most of the expression variance in yeast (Yvert et al., 2003), mouse (Schadt et al., 2003) and human (Whitney et al., 2003; Morley et al., 2004). These studies suggest that only a minority of effects are likely to be operating in *cis*, and that a number of regions show strong linkage to expression of many genes, possibly representing hotspots of transcriptional regulation (Morley et al., 2004). These polymorphisms could participate to diseases, as some diseases are caused by *trans*-acting polymorphisms, including for example familial Alzheimer's disease and Rett syndrome (Gabellini et al., 2003).

At present, most approaches used to measure differential gene expression, such as single-base extension genotyping of RT-PCR products (Cowles et al., 2002; Yan et al., 2002) or hybridisation of mRNA to Affymetrix HuSNP genotyping arrays (Lo et al., 2003), rely on the presence of an exonic or intronic SNP in the transcripts. This restriction is alleviated by a recent approach called 'haploChip' which allows indirect measure of gene expression in vivo, namely phosphorylated Pol II loading on DNA in living cells (Knight et al., 2003). By allowing the analysis of any SNP within 2 Kb 5' or 3' to the gene, including promoter and 3'UTR SNPs, this method considerably expands the number of haplotypes that can be interrogated (Knight et al., 2003). It is noteworthy that elucidation of allele-specific gene expression using SNPs to discriminate transcribed RNA serves to identify functionally important haplotypes, rather than specific variants, unless sufficient complexity of the underlying haplotypic structure can be resolved to fine map observed differences down to a specific region or individual polymorphism (Knight, 2004). Furthermore, consequences of regulatory polymorphisms are highly context dependent on other genomic sites, both within and outside the locus, on the cell type, and on environmental conditions and exogenous stimuli (Knight, 2005).

1.6 Gene transcription

Gene expression is the fundamental process governing protein content and turn-over in cells of unicellular and multi-cellular organisms whereby the genetic information stored in the relatively inactive DNA is transferred into active molecules (Svejstrup, 2004). Gene expression follows three main steps: transcription, translation, and post-translation modification. Transcription is the step in which mRNA molecules are produced according to the information coded in a gene and occurs in the nucleus (Proudfoot et al., 2002). Three RNA polymerases function in eukaryotes, and transcription of protein-coding genes is mediated by DNA-dependent RNA-polymerases II (RNApolII). In the nucleus, DNA is coiled around histones, forming primary elements of packaged DNA, or nucleosomes, themselves further organised into chromatin (Dillon, 2004; Verschure, 2004). Access to DNA by RNApolII requires chromatin remodelling, via histone modifications (Workman and Kingston, 1998; Cho et al., 2004; Li et al., 2004c). RNApolII binds to general transcription factors to form the pre-initiation complex which binds DNA, allowing RNApolII to start initiation (Forget et al., 1997; Muller and Tora, 2004; Cramer, 2004). It has recently emerged that, after initiation, transcription and mRNA processing do not occur in the series of distinct events as thought previously but instead happen simultaneously (Orphanides and Reinberg, 2002). Initiation of transcription is followed by elongation, which is accompanied by maturation of the transcript as it is produced (Orphanides and Reinberg, 2002; Proudfoot, 2004; Shilatifard, 2004), involving its capping, for protection from nucleases and then transport to cytoplasm and translation, its splicing, and its termination by poly-adenylation (Stamm et al., 2005; Proudfoot et al., 2002). RNA performs various structural and catalytic functions, but its main role, however, is to be transcribed into proteins. This requires its transport into the cytoplasm (Reed and Hurt, 2002), where it is transcribed into a chain of amino acids by ribosomes (Ramakrishnan, 2002; Dever, 2002), which is then folded into a protein (Daggett and Fersht, 2003). Most proteins are then modified so that they can carry out structural and metabolic processes required for the cells to function properly (Morrison et al., 2002; Aebersold and Goodlett, 2001). Transcription affects all biological processes from development to response to environmental factors. Although gene expression is also

controlled at the level of translation and beyond, regulation of gene expression is mostly achieved by transcription mechanisms (White, 2001). Furthermore, initiation of transcription is one of the most important levels of gene expression regulation (White, 2001).

Regulatory elements can be divided into two major groups, depending on whether they are recognised at the DNA or RNA level. The first group includes elements important at the DNA level: promoters, enhancers, locus control regions (LCRs) as well as matrix attachment regions/scaffold attached regions (MAR/SARs) (Fessele et al., 2002; Frisch et al., 2002). The second group is composed of elements recognised at the RNA level and involved in post-transcriptional processes, namely *cis*-elements affecting regulation of processing, transport, translation and stability of mRNAs (Reed and Hurt, 2002; Ramakrishnan, 2002; Dever, 2002). Transcription is regulated by elements in the first group and its initiation requires transcription factors (TFs), which recognise and bind to short motifs on the DNA in the vicinity of the gene: Transcription Factor Binding Sites (TFBSs).

1.6.1 Promoter elements

TFBSs are typically 5-25 bp long degenerate sequence motifs (Stormo, 2000a). The binding of TFs to TFBSs guides the binding of the RNA-polymerase to the DNA (Stormo, 2000a; Muller and Tora, 2004; Cho et al., 2004; Cramer, 2004). The combination of short sequence elements to which the RNA polymerase binds in order to initiate transcription of a gene is referred to as promoter (Smale and Kadonaga, 2003). The formation of a complex of proteins containing general transcription factors and the RNA-polymerase II, the Pre-Initiation Complex (PIC), and its binding to the TATA box located 25 bp upstream of the Transcription Start Site (TSS) in Metazoans, activates RNA-polymerase and triggers production of mRNA (Kadonaga, 2004). There is growing evidence that integration of signals at the promoter relies on the specificity and flexibility of the promoter, which in turn are mediated by the organisation of these elements, rather than on an individual element (Fessele et al., 2002; Dvir, 2002).

Another important regulatory element is the Initiator region (Inr), a loosely conserved region around the TSS. It is one of the determinants of the strength of the promoter. In the absence of a TATA box, the Inr can substitute for it and determine the position of the TSS. The TATA box and Inr are interchangeable to some extent. Some proteins can bind to the Inr. These proteins may have a role in the specificity of transcription initiation. In addition, only half of the promoters contain a TATA-box. Most of the promoters without a TATA box contain a loosely conserved sequence element resembling CCAAT, recognised by the CCAAT/Enhancer-binding protein (Kadonaga, 2004).

Another important feature of promoters is the presence of CpG islands close to or overlapping them. CpG islands indicate a lack of methylation at CpG doublets and correlate with the open chromatin configuration. Overall, CpGs occur at less than one fourth of the frequency expected from the C+G content in the human genome. The reason for their rarity seems to be the frequent mutation of the methylated C into a T in most regions of the genome that are not transcriptionally active (Strathdee et al., 2004; Caiafa and Zampieri, 2005).

1.6.2 Non-promoter regulatory elements

The promoter located immediately upstream of the TSS, together with the PIC and specific regulatory elements control the location of the initiation of transcription (White, 2001; Muller and Tora, 2004; Kadonaga, 2004). The latter also determine the specificity of gene expression at the tissue and developmental levels, or any biochemical context (Kadonaga, 2004). Some of these regulatory regions contain several binding sites for specific TFs and have a synergetic action. These regions are therefore referred to as activators or repressors. Enhancers are activators that contain several TFBSs, can be located up to tens of kilobases from the TSS, and are functional independent on their orientation and to some extent their location (Kadonaga, 2004; Lettice et al., 2002). The inducibility of an activator depends upon the sequence of the promoter, and upon the capability of a TF to bind DNA and interact with other TFs (Muller and Tora, 2004; Cramer, 2004; Kadonaga, 2004). In addition, it is not yet well understood how the differences in the promoter sequence translate into differences in receptivity to specific TFs. The PIC composition and the

alternative splicing of some proteins associated with TATA-box-binding proteins may however be involved (Kadonaga, 2004).

Gene expression is also regulated by the three-dimensional structure of DNA. Three structural factors appear to be involved in the sequence specificity in the TATA box: the structure of the DNA bound by TATA-box-binding proteins, the competition between histones and TFs in gene activation or repression, and the existence of TFs that reshape the DNA to bring distant sites to proximity (Muller and Tora, 2004; Dillon, 2004; Cramer, 2004).

1.6.3 Evolution of transcriptional systems

Several studies have reported conservation of regulatory sequences across species, as distantly related as mouse and fish. For example, divergence of axial morphologies has been shown to involve variations in regulatory regions of developmental control genes, such as *HOXB4* and *HOXC8*. An early study showed that regulatory regions for the murine *Hoxb4* gene could be identified by comparing its non-coding sequence to that of the puffer fish *Takifugu rubripes* (Aparicio et al., 1995), despite the fact that ancestors of these two species diverged over 400 millions years ago (Powers, 1991). The authors first identified *Hoxb4* in *T. rubripes*, and then aligned the non-coding sequences of the gene sequences from the two species. Three highly conserved sequences were identified and were localised in regions of the mouse sequence known to contain enhancers. One was intronic and controlled expression in the mesoderm, ectoderm and in the central and peripheric nervous system, and two were in the sequence downstream of the gene and controlled the expression in the ventral neural tube, with a clear anterior boundary. The function of these conserved sequences was tested using gene reporter constructs in transgenic mice. The first region, 92 bp, was deleted from the enhancer and tested in transgenic embryos, which lost the major domains of expression in the mesoderm and ectoderm, but retained expression in the neural tube. This indicated that the enhancer could still function at multiple sites, and more importantly, that this 92-bp sequence was crucial for expression in the mesoderm, and central and peripheric nervous system. The function of each of the remaining two conserved sequences was tested using constructs including the corresponding fish sequence. The second region, 250 bp, did

not show consistent pattern of expression, in contrast to the third region, 300 bp, which mirrored the spatial boundary of expression observed in mouse, identifying a neural element that governs a subset of *Hoxb4* expression conserved in mammals and fish (Aparicio et al., 1995).

Using a similar approach, based on several mammalian species, another study identified a even smaller region in the early enhancer of the *HOXC8* gene showing that new functions or expression patterns can arise by independent loss or gain of regulatory elements (Shashikant et al., 1998). In one of the five known *cis*-acting elements in the early enhancer, which are almost invariant in Mammals, the authors identified a 4-bp deletion in the baleen whales. Transgenic mouse embryos carrying a construct with the whale enhancer showed patterns of expression different from that observed in wild-type mice: a loss of expression in the posterior mesoderm, and a gain of expression in the neural tube in more posterior areas. The authors therefore suggested that the expression of *Hoxc8* is different in whales than in other mammals analysed, and that this may impact on variation in axial morphology (Shashikant et al., 1998). Follow-up studies further characterised existing elements and identified novel elements in the early enhancer, and extended the analysis to fish, supporting further the role of *Hoxc8* in axial morphology in vertebrates (Anand et al., 2003).

Some novel functions and expression patterns have also been caused by mutations in genes encoding TFs (Veraksa et al., 2000; Hsia and McGinnis, 2003; Levine and Tjian, 2003), resulting in additional function (Galant and Carroll, 2002; Ronshaugen et al., 2002), or completely new roles via alteration of interaction with cofactors (Alonso et al., 2001; Brown et al., 2001; Lohr et al., 2001). Mechanisms of the evolution of gene transcription, such as those concerning the organisation of promoters or the binding of TFs to DNA, and especially its specificity, are not well described (Stone and Wray, 2001; Rockman and Wray, 2002; Wray et al., 2003). TFBSs can be highly degenerate and initiation of transcription is dependent upon modules of such sites, rarely upon individual sites (Dvir, 2002; Fessele et al., 2002; Smale and Kadonaga, 2003; Kadonaga, 2004). The degeneracy of TFBS sequences appears to be beneficial because it allows promoter flexibility, and would therefore have been selected through evolution (Stormo, 2000a; Stone and Wray, 2001; Bulyk, 2003; Rockman et al., 2003). The small size of TFBSs, the weak sequence

requirement for TF binding and the modular organisation of TFBSs clusters suggest that: i) mutations may allow for the presence of conservative substitutions in TFBSs, and ii) mutations may eliminate existing TFBS and generate new ones resulting in new expression patterns or functions (Shashikant et al., 1998; Stone and Wray, 2001; Dermitzakis and Clark, 2002; Ludwig, 2002).

To test whether the function of promoters in mammals is conserved despite sequence degeneracy as reported in the fruit fly, Dermitzakis and Clark (2002) identified genes with a promoter containing TFBSs where the binding ability and function had been experimentally verified from the literature. They then performed pairwise comparison of the sequence of the 51 identified human genes to that of homologous sequences in macaque, mouse and rat. Results of the human-macaque comparisons revealed a slow accumulation of substitutions in TFBSs, and showed that these substitutions were not spread evenly across TFBSs but instead concentrated in a few sites only. This was also observed later in comparison of human, mouse and dog sequences (Dermitzakis et al., 2004). The heterogeneous rate of substitution observed among TFBSs could be explained by higher flexibility of the binding ability of some TFs or by selective constraints being more relaxed in some binding sites than others (Dermitzakis and Clark, 2002). Comparisons of human and mouse sequences for 43 gene pairs revealed that the divergence of TFBSs was lower than the average synonymous divergence, but higher than the non-synonymous divergence. Moreover, it was not correlated with the divergence of the amino acid sequences, and therefore could not be explained by rapid overall gene divergence. Results from comparisons of human and macaque, and human and rodent, also suggest that selective constraints on TFBSs are sometimes independent in the two lineages. Finally, comparison of the function of 64 TFBSs in 20 genes, representative of the entire set, in human and rodent showed that TFBSs diverged substantially; 33 were functional in both species, 14 in human only, and 17 in rodents only, leading the authors to estimate that between 32% and 40% of human functional TFBSs are not functional in rodents.

These findings are relevant to the understanding of disease mechanisms, as rare mutations in highly conserved TFBSs may have a strong effect and cause monogenic illnesses, while mutations in less conserved TFBSs, missed by

comparisons between distant species, may have smaller effects, be common and implicated in complex diseases (Dermitzakis and Clark, 2002). This is supported by accumulating evidence showing that polymorphisms in promoter sequences account for a substantial amount of genetic and phenotypic variation observed in populations (Wray et al., 2003; Levine and Tjian, 2003). In addition, selective pressures acting upon non-coding sequences have been shown to vary between loci and clades (Wray et al., 2003; Levine and Tjian, 2003; Rockman et al., 2003). Alignments of sequences from human and vertebrates show that regulatory regions are conserved to a certain degree and comparative genomics has been used successfully to identify regulatory regions (described below). The following section summarises the two types of computational methods used to predict the genes themselves, and describes the methods used to predict *cis*-elements involved in gene transcription.

1.7 Prediction of regulatory regions

To fully understand gene transcription requires the identification of all genes involved and the unravelling of their function. Reciprocally, to understand the function of genes, it is necessary to identify sequence elements that control their expression. Identification of genes and regions responsible for their regulation is one of the main goals of large-scale genome sequencing projects, which now determine hundreds of megabases each year. The most recent and significant advances are the near completion of the assembly of the following vertebrate genomes: human (International Human Genome Sequencing Consortium, 2001), mouse (Waterston et al., 2002), rat (Rat Genome Sequencing Project Consortium, 2004), dog (Kirkness et al., 2003) and chicken (Hillier et al., 2004). One of the major challenges is to extract information from the huge amount of data to detect genes and decipher their function. This is particularly difficult in eukaryotes, even more so in higher eukaryotes, because most of the genomic sequence is non-coding.

1.7.1 Gene prediction

Gene prediction aims to correctly define the structure of a gene, and when a sequence contains several genes it aims to partition a set of predicted exons among these genes. This problem can be approached by two types of techniques, based on

statistical methods or on homology searches. The first type of technique aim to predict the overall gene structure by identifying coding regions using statistical means and pattern recognition for sites for transcription, splicing and translational control (Burset and Guigo, 1996; Stormo, 2000b; Brent and Guigo, 2004). The main problem these programs face is the confusion of internal exons with initial and terminal exons, which is inherent to the difficult identification of the signals for start and stop of transcription and translation and location of the splicing sites (Burset and Guigo, 1996; Stormo, 2000b; Brent and Guigo, 2004). The second type of technique relies on homology searches for close homologs in the databases to one of the genes in the sequence under analysis. The main drawback of this approach is its dependency upon the presence of homologs in the databases whereas novel genes by definition are not related to known genes closely enough so that homology can be detected. In addition, when a homolog is identified, these homologous genes may differ due to insertion and deletion (Stormo, 2000b; Morgenstern et al., 2002; Brent and Guigo, 2004). Therefore, an important goal of gene prediction is to design means of recognising the beginning and end of genes (Davuluri et al., 2001; Down and Hubbard, 2002).

1.7.2 Prediction of *Cis*-elements in promoters

Although DNA conformation plays an important role in gene transcription initiation, little data on structural aspects of transcription initiation is available (Workman and Kingston, 1998; Cho et al., 2004; Li et al., 2004c; Brent and Guigo, 2004). For this reason, most promoter prediction programs are based on the specificity of protein-DNA contacts at the sequence level (Werner, 2000). Initiation of the transcription of a gene depends on the co-ordinate binding of many proteins to the promoter and sometimes to enhancers (Kadonaga, 2004). The combination of different binding sites determines the particular expression context of a gene (Kadonaga, 2004; Smale and Kadonaga, 2003; Dvir, 2002). Comprehensive annotation of regulatory sites therefore requires identification of individual sites and their functional combination (Gailus-Durner et al., 2001; Bulyk, 2003; Werner, 2003a; Werner, 2003b; Werner et al., 2003; Pavesi et al., 2004)

1.7.2.1 Identification of known Transcription Factor Binding Sites

To identify known TFBSs in a sequence, one can use data obtained experimentally and publicly available, and determine which known regulatory element is present in the sequence. DNA-protein binding specificity of a site can be modelled in two ways: by a consensus sequence or by a position weight matrix (PWM) (Werner, 2000; Stormo, 2000a; Bulyk, 2003). A consensus sequence is created from a collection of known TFBS by determining the most frequent base at each position within a site. Matching the consensus sequence to the sequence under scrutiny identifies occurrences of the motif. Mismatches must therefore be allowed to represent variation of the pattern. The main drawback of this approach is thus its inherent loss of information (Werner, 2000). A weight matrix is computed based on a collection of TFBS by measuring the frequency of each element at each position of the site. Any putative site is assigned a score computed from the sum of the matrix values for that sequence. The first limitation of PWM is that a threshold must be defined to select predicted motifs. The second limitation is that different positions are assumed to make independent contributions to DNA binding (Werner, 2000; Bulyk, 2003).

Four main databases exist that contain information on known promoters: Eukaryote Promoter Database (Perier et al., 2000), TRANSFAC (Knuppel et al., 1994; Matys et al., 2003), TRRD (Heinemeyer et al., 1998) and JASPAR (Sandelin et al., 2004). There are several computational tools for the prediction of TFBSs based on consensus sequences (EMBOSS tool tfscan, http://www.emboss.org) or based on sets of position weight matrices, e.g. MatInspector (Quandt et al., 1995) and MATCH (Kel et al., 2003). To reduce the large sequence space to search and increase chances to identify location of genuine functional sites, numerous studies have focused on sequences of genes that show highly similar expression patterns (Roth et al., 1998; Tavazoie et al., 1999; Hughes et al., 2000; Bussemaker et al., 2001; Werner, 2003b; Werner et al., 2003). It is noteworthy that the accuracy of the prediction is directly influenced by the quality of the library of weight matrices used. In particular, many of the sets used to calculate the PWM are small and predictions may therefore not be accurate (Roulet et al., 2000).

As for gene prediction, two types of approach exist to predict regulatory elements. The first type integrates biological knowledge to delineate a set of definitions to determine regulatory elements (Ohler and Niemann, 2001; Stormo, 2000a; Bulyk, 2003; Pavesi et al., 2004). The second type of approach is based on the comparative analysis of homologous sequences, also called phylogenetic footprinting (Tagle et al., 1988; Dubchak and Frazer, 2003; Frazer et al., 2003). These methods have been increasingly used since the genomes of several vertebrate model organisms have been sequenced (Thomas et al., 2003).

1.7.2.2 Identification of putative novel regulatory sequence motifs

Methods used to detect motifs in sequences rely on the computation of the expected frequency of motifs in these sequences in order to identify motifs that occur more frequently than expected by chance alone. Two types of approach exist to detect motifs over-represented in a set of sequences, which rely on sequence alignment or enumeration of motifs.

The underlying principle of methods based on sequence alignment is to align sequences of interest to identify elements in several or all sequences. Two methods can achieve this goal: Gibbs sampling, implemented in AlignACE (Roth et al., 1998) and expectation-maximisation in MEME (Bailey and Elkan, 1994). The typical output is a model for each motif identified, usually a PWM.

Enumerative methods rely on the frequency of all nucleotide oligomers of various size in the genome of interest to identify motifs that are over-represented in the searched sequences. This approach is implemented in the programs Oligo-analysis, which detects simple oligomers (van Helden et al., 1998), and Dyad-analysis, which detects homo- or hetero-dimer motifs separated by a fixed or variable spacer length (van Helden et al., 2000). The typical output of these two programs is a list of over-represented oligomers, represented in the form of a consensus sequence (van Helden et al., 2000).

Methods based on sequence alignment have the advantages over enumerative methods of finding potentially longer motifs and delivering a weight matrix as a comprehensive model for a motif, which is more flexible than the consensus sequence generated by enumerative methods. In addition, parts of this consensus

sequence may come from different sequences. The main advantage of enumerative methods is that they search all possible oligomers (Stormo, 2000a; Werner, 2000).

Due to the small size and relative sequence degeneracy of TFBSs, prediction of such motifs generates many false positives (Werner, 2000; Stormo, 2000a). The problem can be addressed by considering only those located in sequences conserved in another organism.

1.8 Comparative genomics

1.8.1 Human-mouse sequence comparison

The underlying assumption of the comparative genomics approach, or phylogenetic footprinting, is that sequences conserved through evolutionary time have not diverged because they have a necessary function (Tagle et al., 1988). Early examples of comparative sequence analysis in mammals include the study of globin genes, in the context of thalassemia, in human and rabbit (Hardison et al., 1991), with a dot-plot software to visualise the pairwise alignment along the sequence annotation (Hardison et al., 1991; Schwartz et al., 1991), and squirrel (Gumucio et al., 1996; Hardison et al., 1997), and further developed in a database, Hbvar (Hardison et al., 2002; Patrinos et al., 2004). Comparative genomics of the human and mouse sequences has been successfully used to identify coding sequences (Batzoglou et al., 2000), and integrated in gene prediction programs, such as SLAM (Alexandersson et al., 2003) and TWINSCAN (Korf et al., 2001; Flicek et al., 2003)

Comparative sequence analysis of human and mouse sequences has been applied to the evolution of chromosome structure and organisation (Dehal et al., 2001), as well as specific regions of a chromosome involved in human diseases, such as the Wolfram syndrome, leading to the identification of novel genes (DeSilva et al., 2002), and regulatory regions (Wasserman et al., 2000; Pennacchio and Rubin, 2001; Dermitzakis and Clark, 2002). For example, Loots and co-authors (2000) reported that comparison between sequences of approximately 1 Mb from human and mouse allowed the identification of regions of at least 100 bp with percentage of identity higher than 70% (Loots et al., 2000). Analysis of 15 of these sequences showed that 70% were conserved in mammals and characterisation of the largest conserved non-

coding sequence in transgenic mice showed that it controlled the expression of three genes (Loots et al., 2000). In another study, an analysis of individual TFBSs, Wasserman and colleagues (2000) showed that 74 out of 75 experimentally defined binding sites were located in the 19% of the human sequences that are the most conserved in the mouse orthologous sequence (Wasserman et al., 2000). Similarly, by combining searches of TFBSs in TRANSFAC with comparative analysis of sequences from human and mouse, Levy and Hannenhalli (2002) found that although only half of the 481 TFBSs were found in non-coding sequences conserved in mouse, the predictive power was increased by a up to three-fold factor using evolutionary conservation (Levy and Hannenhalli, 2002). Another more elaborate approach attempted to distinguish between functional and neutrally evolving non-coding sequences, based on sequences features (Elnitski et al., 2003). Percentage identity plots of aligned sequences and other tools have also enabled clearer visualisation, such 'PipTools' (Schwartz et al., 2000; Elnitski et al., 2002), Alfresco (Jareborg and Durbin, 2000) and VISTA (Mayor et al., 2000; Loots et al., 2002).

The main difficulty in the comparative analysis of long non-coding sequences is the detection of weak similarities. Several programs have been developed to solve this problem: MUMmer (Delcher et al., 2002), Bayes Block Aligner (Zhu et al., 1998), BLASTZ (Schwartz et al., 2003b), and LAGAN (Brudno et al., 2003). For example, Bayes Block Aligner identifies colinear conserved blocks separated by non-conserved sequences of certain length and identifies blocks with identity > 60%. This method was used to show that the 100 bp upstream of the gene start site was conserved in 70% of the gene pairs (Jareborg et al., 1999). LAGAN identifies conserved sequences by joining together very sensitive local alignments constructed based on multiple inexact words (Brudno et al., 2003).

Bioinformatics approaches are rapidly evolving in pace with the increase in genome sequence and annotation. The work presented in chapter 4 aimed to annotate regulatory elements of genes in the 4p region linked to psychoses. The protocol presented was designed in 2002 and combined the three types of computational approach for predicting regulatory elements: comparative genomics, prediction of known TFBSs and identification of over-represented motifs. This protocol was tested on a set of genes for which functional TFBSs were available publicly (Dermitzakis

and Clark, 2002). First, human repeat-masked non-coding upstream regions of the test genes were searched against upstream regions of mouse genes. Conserved sequences were then scanned for known TFBSs in TRANSFAC, and over-presented motifs were predicted using programs implementing the enumerative oligo-counting approach and programs based on sequence alignment. Accuracy of the protocol was assessed by determining its ability to predict motifs that match TFBS. In view of the results of this work, a purely comparative genomics approach was used to identify non-coding regions conserved in several vertebrates species.

1.8.2 Multiple species comparative genomics

It has recently been suggested that 5% of the mammalian genome is under selective constraint (Waterston et al., 2002), while a substantial proportion of the conservation between human and rodents includes neutral sequences that may not have diverged yet (Rat Genome Sequencing Project Consortium, 2004). Comparing genomic sequences from more than two species has the advantage over studying two species of reducing the amount of conserved neutral sequences, as divergence eliminates these sequences differently in each species and lineage (Duret and Bucher, 1997; Thomas et al., 2003). Various multiple alignments programs have recently been published, such as MultiPipMaker (Schwartz et al., 2003a), MAVID (Bray and Pachter, 2004), and MLAGAN (Brudno et al., 2003), which should allow for a sensitive detection of conserved functional non-coding sequences (Frazer et al., 2003; Nobrega and Pennacchio, 2004). Methods for scoring sequence conservation in multiple alignments have ranged from simple schemes (Tagle et al., 1988; Nobrega et al., 2003; Chapman et al., 2004) to more sophisticated ones (Sumiyama et al., 2001; Elnitski et al., 2003) and performed with sequences from many species (Margulies et al., 2003; Thomas et al., 2003; Chapman et al., 2004).

The human '*dachshund homolog 1*' (*DACH1*) gene, for example, is located in a gene desert and is involved in the development of brain, limbs and sensory organs, and seems to lack upstream control regions (Caubit et al., 1999; Davis et al., 1999; Machon et al., 2002). Phylogenetic footprinting was applied to human and mouse sequences of approximately 2.6 Mb to identify enhancers for the human *DACH1*

gene (Nobrega et al., 2003). Out of approximately 1000 sequences at least 100 bp conserved in human and mouse, 32 remained after including sequences from frog, zebrafish and two species of pufferfish. Out of these, nine were tested in mouse transgenic reporter gene assays, and seven reproduced the expression pattern of the endogenous DACH1. The linear relationship of the newly discovered regulatory elements was also conserved in all species tested (Nobrega et al., 2003). In addition, Gottgens and colleagues (2000) identified a neural enhancer of the *T-cell acute lymphocyte leukemia 1* (*TAL1*) gene which encodes a transcription factor involved in hemopoiesis and vasculogenesis, by i) comparing human, mouse and chicken sequences, ii) identifying a conserved sequence with no previously known function, and iii) showing in a transgenic Xenopus reporter assay that this sequence was functional (Gottgens et al., 2000). Follow-up studies identified additional important regulatory regions using sequences from zebrafish and pufferfish (Gottgens et al., 2002), and dog and rat, showing the advantage of multiple species over pairwise sequences analysis (Chapman et al., 2004). Furthermore, a recent study using alignments of sequences of the CFTR region from 12 vertebrate species showed that human functional sequences are best identified by considering a wide range of mammalian sequences (Thomas et al., 2003). However, for most genes in vertebrates, the readily available genomic sequence is restricted to human (International Human Genome Sequencing Consortium, 2001), dog (*canis vulgaris*) (Kirkness et al., 2003), mouse (*Mus musculus*) (Waterston et al., 2002), rat (*Rattus norvegicus*) (Rat Genome Sequencing Project Consortium, 2004), chicken (*Gallus gallus*) (Hillier et al., 2004), zebrafish (*Danio Rerio*) (http://www.zfin.org), and two species of pufferfish; *T. rubripes* (Aparicio et al., 2002) and *Tetraodon nigroviridis* (Crollius et al., 2000). The accuracy of multiple species comparative genomics has usually been assessed in terms of its ability to detect exons (Margulies et al., 2003) since a recurring problem in the prediction of non-protein-coding sequences is the scarcity of such sequences with a experimentally verified function to act as positive controls.

1.9 Aims

The aims of this research were to develop and use computational tools to facilitate the characterisation of candidate genes in the 4p15-16 region linked to psychosis. Specifically, the aims were:

- the development of a computational tool to assist ongoing laboratory work by allowing association data management and analysis,
- the *in silico* determination of putative functional non-coding sequences,
- the identification of SNPs located in these regions which are candidate markers for large-scale case-control allelic association studies.

CHAPTER 2

MATERIAL AND METHODS

2 MATERIAL AND METHODS

2.1 Allelic association data

Data were managed in an ACeDB database (http://www.acedb.org). The CGI script was written in Perl and uses the CGI.pm Perl5 library (http://stein.cshl.org/WWW/software/CGI/cgi_docs.html).

2.2 Combination of human-mouse comparative genomics with sequence motif prediction

2.2.1 Protocol

A schema describing the protocol is shown in figure 2.1. Repeats in human sequences containing known functional *cis*- regulatory elements and upstream sequences of known genes in mouse were masked using RepeatMasker (Smit and Green, unpublished). Repeat-masked human sequences were searched against repeat-masked mouse sequences. Human sequences conserved in mouse were selected based on the percentage of identity and length of the alignment. Non-conserved functional *cis*-elements were not considered further. Known Transcription Factor Binding Sites (TFBS) were predicted on the selected conserved regions of the human sequences by searching TRANSFAC by using the tfscan program (http://www.emboss.org). Selected conserved regions of the human sequences matched by at least four mouse sequences were used together with the matching mouse regions to predict novel motifs using four motif prediction programs. Location of the conserved functional *cis*-elements was then compared to that of predicted motifs to determine which of these *cis*-elements were identified by predicted motifs. Accuracy was assessed for each motif prediction program as well as for a series of flexible combinations of these programs.

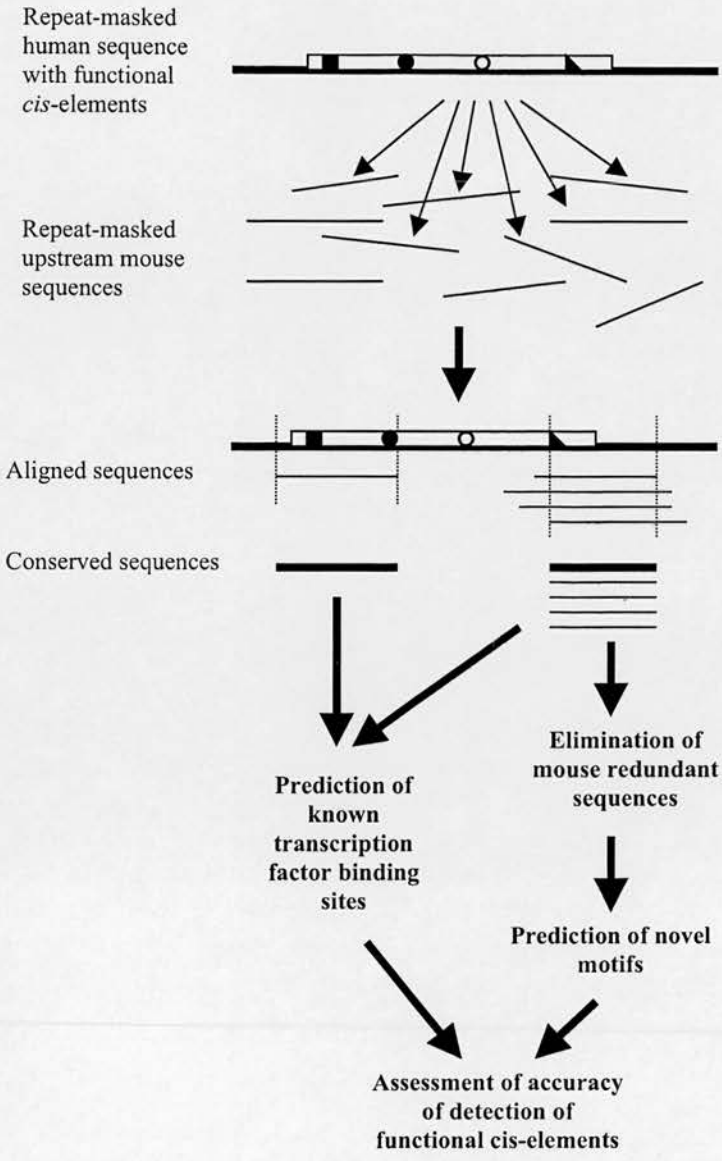


Figure 2.1: Prediction of regulatory motifs using human-mouse comparative genomics and prediction of known and novel sequence motifs. The repeat-masked human sequence containing functional *cis*- regulatory elements, that is a region (open rectangle) and sequence motifs (square, disc, circle and triangle within the open rectangle), is searched against sequences upstream of known genes in mouse. Hits are mapped onto the human sequences to identify conserved sequences. Conserved sequences are searched for known Transcription Factor Binding Sites from the TRANSFAC database. The sub set of conserved sequences matched by at least four hits is used to predict novel motifs. Accuracy of detection is then assessed by comparing the location of the predicted motifs to that of the conserved known functional *cis*-elements.

2.2.2 Sequence data

2.2.2.1 Human sequences

Human genes were chosen from sets available publicly (Dermitzakis and Clark, 2002). These genes were selected on the basis of experimental identification of functional regions in promoter sequences. These functional regions were derived based on direct experimental confirmation of binding ability (footprinting, gel shift assay) and function (promoter deletion experiments, directed mutagenesis, expression of reporter genes) (Dermitzakis and Clark, 2002). Two types of functional *cis*-elements were considered: i) 'regulatory regions', which are required for the activity of the promoter, and ii) 'regulatory motifs', which are shorter sequences, such as TFBS. The list of 28 genes selected is shown in Table 4.1.

2.2.2.2 Mouse sequences

Regions spanning 2 Kb upstream of the translation start site of known genes and genes predicted by *ab initio* gene prediction programs supported by similarity to Expressed Sequence Tags and/or protein sequences, were retrieved from the Ensembl web site using the EnsMart tool available on this site (Clamp et al., 2003). Flanking sequences were retrieved for 13 591 genes from the mouse genome assembly 12.3.1 (as in March 2003).

2.2.3 Conservation of human sequences in mouse

Human sequences were each searched against the set of upstream regions of genes in the mouse genome using LAGAN (Brudno et al., 2003) with the default parameters. Human sequences longer than 50 bp and sharing more than 65% identity with mouse sequences were selected for further analysis. These sequences are hereafter referred to as 'CSSs' (for 'Conserved Sub-Sequence with single mouse sequence').

Evolutionary conservation of each human functional *cis*-element was analysed by comparing the location of the *cis*-element to that of human conserved sequences (i.e. 'CSSs'). Conservation was defined based on the proportion of the

length of the *cis*-element overlapped by the selected human conserved sequence. Each *cis*-element which was overlapped by a conserved sequence by at least 50% of its length was considered conserved between human and mouse.

2.2.4 Prediction of known and novel motifs

2.2.4.1 Prediction of known transcription factor binding sites

Known TFBSs were identified in conserved human sequences by searching the TRANSFAC 'vertebrate' database using the EMBOSS tool tfscan (http://www.emboss.org), not allowing mismatches. Predicted motifs known in human, mouse and rat, and at least 6 bp long were considered in the analysis.

2.2.4.2 Prediction of novel motifs

Conserved human sequences were separated into two classes according to the number of regions they matched in the set of mouse sequences (Figure 2.1). The first class contained conserved human sequences matched by less than four mouse sequences. The second class contained conserved human sequences matched by at least four mouse sequences. For each human sequence in the second class, a set of sequences was created that contained the human conserved region and mouse regions that matched this human region. In each set of human and mouse sequences, redundant mouse sequences were identified as sharing a percentage of identity higher than 80%. Such redundant mouse sequences were removed from each set. Sets of sequences were thereby created that contained the human conserved sequence and the selected mouse sequences matching the human region. These sets of sequences are referred to as 'CSSm' (for 'Conserved Sub-Sequence with multiple sequences') hereafter. These sets were used to predict novel motifs using the following programs: AlignACE, based on Gibbs sampling (Hughes et al., 2000), MEME, based on Expectation-Maximisation (Bailey and Elkan, 1994), and Oligo-analysis (van Helden et al., 1998) and Dyad-analysis (van Helden et al., 2000) which both implement the oligo-counting method. MEME was run in 'oops' mode. Oligo-analysis was set to predict motifs of 6 bp. Dyad-analysis was set to identify motif with a spacer of length

varying from 0 to 20. Selection thresholds used for these four programs were: i) a minimum score of 10 for motifs predicted by AlignACE, ii) an e-value less than 10^{-3} for motifs predicted by MEME, and iii) a score higher than 3 for motifs predicted by Oligo- and Dyad-analysis.

2.2.5 Detection of conserved functional *cis*-elements

2.2.5.1 Individual prediction methods

Known and novel motifs were predicted on sequence regions conserved between human and mouse. Known functional *cis*-elements not conserved between human and mouse were therefore not considered further and only those that were conserved between human and mouse were included in this analysis. Identification of conserved known functional *cis*-elements by known and novel predicted motifs was assessed by comparing the location of these *cis*-elements to that of each motif predicted on the conserved human region containing these *cis*-elements. A conserved element was considered to be identified by a predicted motif if the length of the overlap of the *cis*-element by the predicted motif was greater than 50% of the length of the *cis*-element. Individual prediction methods were compared based on this information.

2.2.5.2 Flexible combinations of motif prediction methods

For each conserved human sequence, regions matched by one or several motifs predicted by one or more prediction programs were identified. Clusters of predicted motifs were defined as a set of overlapping predicted motifs. The region on the conserved human sequence corresponding to each cluster of motifs was defined as the segment overlapped by this cluster. Clusters of overlapping motifs each contained motifs predicted by one or more prediction programs. These clusters were classified based on the number of programs which predicted these overlapping motifs, i.e. clusters made of motifs predicted by only one program, clusters made of motifs predicted by any two programs and so on.

2.2.5.3 Accuracy of detection of known functional *cis*-elements

Each predicted motif was classified as a True Positive (TP) if at least 50% of its length overlapped known functional *cis*-elements, and as a False Positive (FP) otherwise. Each known functional *cis*-element was classified as a False Negative (FN) unless 50% or more of its length was overlapped by predicted motifs. Each known non-functional sequence was classified as True Negative (TN) if less than 50% of its length was overlapped by predicted motifs. Sensitivity (SS) was measured as the ratio of the number of conserved functional *cis*-elements identified to the total number of conserved functional *cis*-elements (SS = TP / (TP + FN)). Specificity (SP) was computed as the percentage of true negatives correctly predicted (SP = TN / (FP + TN)). The positive predictive value (PPV) was computed as the proportion of predicted motifs that identified known functional *cis*-elements (PPV = TP / (TP + FP)).

2.3 Evaluation of scoring schemes for multiple species comparative genomics

2.3.1 Sequence data

Human genes were chosen from a publicly available set of genes known to have one or more experimentally verified regulatory regions (Elnitski et al., 2003) situated within the 2 Kb region upstream of the transcription start site of the gene. For each human gene, its orthologs in mouse, rat, chicken, zebrafish and pufferfish *T. rubripes* (referred to as 'fugu' hereafter) were identified using the Ensembl web site (http://www.ensembl.org and http://pre.ensembl.org/Galllus_gallus, as in April 2004) by curating reciprocal best BLAST hits reported by Ensembl (Clamp et al., 2003) and generated manually by searches using the Ensembl online blast tool. For each set of orthologs, conflicting data, often due to the presence of paralogs and incorrect annotation, was resolved by manually inspecting the extent of synteny conservation. Genomic sequences used in the analysis comprised of the gene sequences and their 2 Kb flanking regions. These sequences and their annotation were obtained from Ensembl (Clamp et al., 2003). Four sets of sequences were studied. First, the 'HMR'

set consisted of each human gene and its orthologs in both mouse and rat ('HMR' stands for 'Human-Mouse-Rat'), and contained 40 such groups of orthologs. Second, the 'HMRC' set consisted of the human genes and their orthologs in both rodents and in chicken ('HMRC' stands for 'Human-Mouse-Rat-Chicken'), and contained 22 groups of orthologs. Third, the 'HMRZF' set consisted of the human genes and their orthologs in both rodents, and in zebrafish and fugu ('HMRZF' stands for 'Human-Mouse-Rat-Zebrafish-Fugu'), and contained 19 groups of orthologs. The fourth set, 'HMRCZF', consisted of the human genes and their orthologs in both rodents, chicken, zebrafish and fugu ('HMRCZF' stands for 'Human-Mouse-Rat-Chicken-Zebrafish-Fugu'), and contained 11 sets of orthologs.

2.3.2 Measure of neutral evolution

For each human gene, rates of synonymous substitution, dS, were computed by maximum-likelihood relative to each non-human species using the program codeml from the PAML package (Yang, 1997) based on multiple alignments of the coding sequence from the species considered in the analysis. Gaps from multiple alignments of protein sequences built using CLUSTALW (Thompson et al., 1994) were propagated to coding DNA sequences.

2.3.3 Multiple alignment of genomic sequences

Multiple alignments of the genomic sequences were built using MLAGAN (Brudno et al., 2003) with repeat-masking by RepeatMasker (Smit and Green, unpublished). MLAGAN is a global multiple alignment tool based on LAGAN which is an anchored global pairwise alignment program that detects the best global alignment by joining an ordered set of local alignments between two sequences. LAGAN uses the CHAOS program to generate local alignments. CHAOS is a highly sensitive method that identifies local alignments based on multiple short inexact words, rather than longer exact words (Brudno et al., 2003).

2.3.4 Scoring systems

2.3.4.1 Principle

Multiple alignments were scanned using a sliding window to compute the pairwise percentages of identity between the human sequence and each of the non-human sequences. These were used to derived a score for each window. Four scoring systems were compared, which are described in the following two paragraphs. Three values were used for the window length: 20 bp, 40 bp, 100 bp, sliding by 10 bp, 20 bp and 20 bp, respectively.

2.3.4.2 Scoring schemes using weighted percentage of identity

The first system was the simple mean of the non-human species' pairwise percentage of identity to the human sequence. The second and third systems involved weighting the pairwise percentage of identity to the human sequence. The second system was based on a simple weight reflecting the relative evolutionary distances between human and the other species, that is 1 for the rodent sequences, 3 for the chicken sequences and 4.5 for the fish sequences, since the evolutionary distance between human and rodent is approximately 100 millions years, that between human and chicken is approximately 300 millions years and that between human and fish is approximately 450 millions years. For the third measure, the weight was the estimate of synonymous substitutions, dS, for each species, or 1 if dS was greater than 1.

2.3.4.3 Scoring scheme based on the rate of neutral evolution

The fourth scoring system used binomial probabilities. Such a scoring system was recently shown to perform well on multiple mammalian sequences (Margulies et al., 2003). This scoring system involves computing the probability of observing the calculated pairwise percentage of identity given the estimate of the rate of neutral evolution, dS, of the gene in the species being compared to human. The score assigned to each window for a given pair of species is the cumulative binomial probability of observing at least as many sites containing identical nucleotides as counted, given dS for the species compared to human. When dS was greater than one, the probability of the two nucleotides at a site being identical was set to 0.25. Otherwise, this probability was the sum of the probability that no substitution

occurred (1-dS) added to the probability that a substitution occurred (dS) multiplied by the probability that the substituted nucleotide is identical to the nucleotide in human (0.25) hence (1-dS)+(0.25*dS), or 1-3dS/4. The score for each clade was computed by averaging over the species in the given clade. For each scoring system the window's final score was obtained by averaging over clades (Margulies et al., 2003). The highest scoring windows were selected until the total length of sequence included in the selected windows reached the proportion of coding sequence in the human sequence analysed. This threshold was therefore specific to each human sequence considered. Conserved segments were then identified by determining the boundaries of the regions covered by the selected windows.

2.3.5 Accuracy of detection of known regulatory regions

To assess the accuracy of the detection of known coding and regulatory regions, positive sites were defined as bases in the coding and regulatory regions and negative sites were the remaining bases. The accuracy of each prediction method was assessed at the nucleotide level by counting the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Sensitivity (SS = TP/(TP+FN)), specificity (SP = TN/(TN+FP)) and positive predictive probability (PPV = TP/(TP+FP)) were then derived. Since the sequences examined are likely to contain unknown regulatory sequences, these estimates should be regarded as conservative.

2.3.6 Sequence conservation in dog, opossum and frog

Whole genome shotgun read sequences for the dog (*Canis familiaris*), gray short-tailed opossum (*Monodelphis domestica*) and western clawed frog (*Xenopus tropicalis*) were obtained from the Ensembl Trace server (http://trace.ensembl.org as in March 2004). Human sequences larger than 40 bp were searched against the sequences of these four species using BLAST (Altschul et al., 1990). Sub-sequences conserved in the dog, opossum and frog were inferred by defining the regions on the query sequences that were matched by at least one shotgun read. The quality of the conservation of the sub-sequences was measured using their score density (score/length) which was computed for each conserved sub-sequence by calculating the mean score-density over the overlapping hits constituting the conserved sub-sequence.

2.4 Comparative analysis of genes in a chromosomal region linked to psychosis

2.4.1 Sequence data

Genes in human chromosome 4p15-16 region linked to psychosis were selected based on the annotation from Ensembl. Only known genes were analysed. These genes corresponded to *ab initio* prediction supported by expression data (Clamp et al., 2003). For each human gene, its orthologs in mouse, rat, chicken, zebrafish and fugu were identified using the Ensembl web site (http://www.ensembl.org and http://pre.ensembl.org/Gallus_gallus, as in April 2004) as described in section 2.3.1. Genomic sequences used in the analysis comprised of the gene sequences and their 10 Kb flanking regions. These sequences and their annotation were obtained from Ensembl (Clamp et al., 2003). Several sets of sequences were considered depending on the species in which orthologs could be identified (see section 2.3.1).

2.4.2 Analysis

Rates of synonymous substitution were computed as described above (section 2.5.2). For each human gene, the sequence and that of the orthologs from the different species were aligned using MLAGAN as described in section 2.3.3. Multiple alignments were then scanned and scored, and highly conserved regions were identified using the scoring scheme based on the probability of the observed level of conservation given the rate of neutral evolution (see section 2.3.4). Accuracy of detection of known coding sequences was assessed in terms of sensitivity, specificity and positive predictive value (as in section 2.3.5).

CHAPTER 3

ALLELIC ASSOCIATION DATA MANAGEMENT

3 ALLELIC ASSOCIATION DATA MANAGEMENT

3.1 Introduction

The large-scale case-control association study of the 4p15-16 region aimed to narrow down the large linkage region and to test specific candidate genes. This approach was expected to include approximately 400 individual DNAs for each disease phenotype and 400 individual DNAs from controls, genotyped at hundreds of SNPs and microsatellites. When the thesis project presented here began, a methodology was being developed in the laboratory by S. LeHellard to cope with the large number of markers and samples to be tested. The protocol tested aimed at genotyping markers in pooled samples of DNA from cases or control individuals, and performing rigorous comparisons of allele frequencies in pooled DNA limiting labour and cost. Markers showing association to the phenotype tested would then be typed in individuals to perform classical case-control association studies. Three genotyping methods were tested by S. LeHellard to type pooled DNA at SNP loci: primer extension followed by allele detection using either dHPLC or mass spectrometry, and SNaPshot (Le Hellard et al., 2002). This chapter provides brief information on the above-mentioned methodology that is necessary to understand the following description of the computational tool developed to manage data generated by association studies based on pooled DNA and individual genotypes.

Inferring allele frequencies from pooled DNA is hindered by experimental difficulties inherent to DNA pooling. Techniques used to distinguish alleles according to the size of the corresponding PCR products generate data in the form of peaks. In the typical output of SNP genotyping, the two alleles are represented by two peaks (figure 3.1a). In a pool, the size of each peak is measured by its area or height and positively correlates with the number of alleles in the sample corresponding to that peak. In an ideal situation, both alleles of a heterozygote are amplified equally (figure 3.1a) and allele frequencies can be derived directly (figure 3.1b). However, due to features inherent to genotyping methods, alleles of a locus are

not amplified with the same efficiency. It follows that the two peaks generated by the genotyping of a heterozygote, which contains one copy of each allele, have a different size (figure 3.1c). The phenomenon is known as 'unequal allele amplification' and needs to be accounted for to infer allele frequencies in pooled DNA (figure 3.1d). The principles of allelic association studies based on pooled DNA are to: i) measure peak heights obtained for a series of heterozygotes, hence for known allele numbers and frequencies, ii) compute the correction factor for unequal allelic amplification, which is the mean ratio of peak heights over the several heterozygotes ('k' in figure 3.1c), iii) correct peak heights obtained for pooled DNA for unequal allelic amplification, iv) derive allele frequencies and numbers in pooled DNA (figure 3.1) and v) use these estimates in the χ^2 test to assess whether alleles at the locus typed are associated to the phenotype tested (Barcellos et al., 1997; Risch and Teng, 1998). For each marker, unequal allele amplification was measured in-house by genotyping several heterozygotes and performing several replicates for each individuals. The ratio of the peak height observed for each allele was then derived. This ratio was found to vary between replicates of individual heterozygotes and between different heterozygote individuals (Le Hellard et al., 2002). Estimating frequencies from peak heights was therefore likely to introduce a non-negligible error. However, the mean peak height ratio for unequal amplification could be accurately measured for any given SNP typed deriving a correction factor, k. This correction factor was used to infer allele frequencies from the reported allele peak heights for SNPs typed in pools, allowing comparison of allele frequencies between pools of DNA (Le Hellard et al., 2002). Genotyping hundreds of markers in DNA pools or hundreds of individual samples generates a large amount of data. This research project therefore required an easily accessible database to store and analyse association data. The next section considers the suitability of available technologies in this context and describes the issues that influenced the design of the tool developed to allow the management of association data in our laboratory.

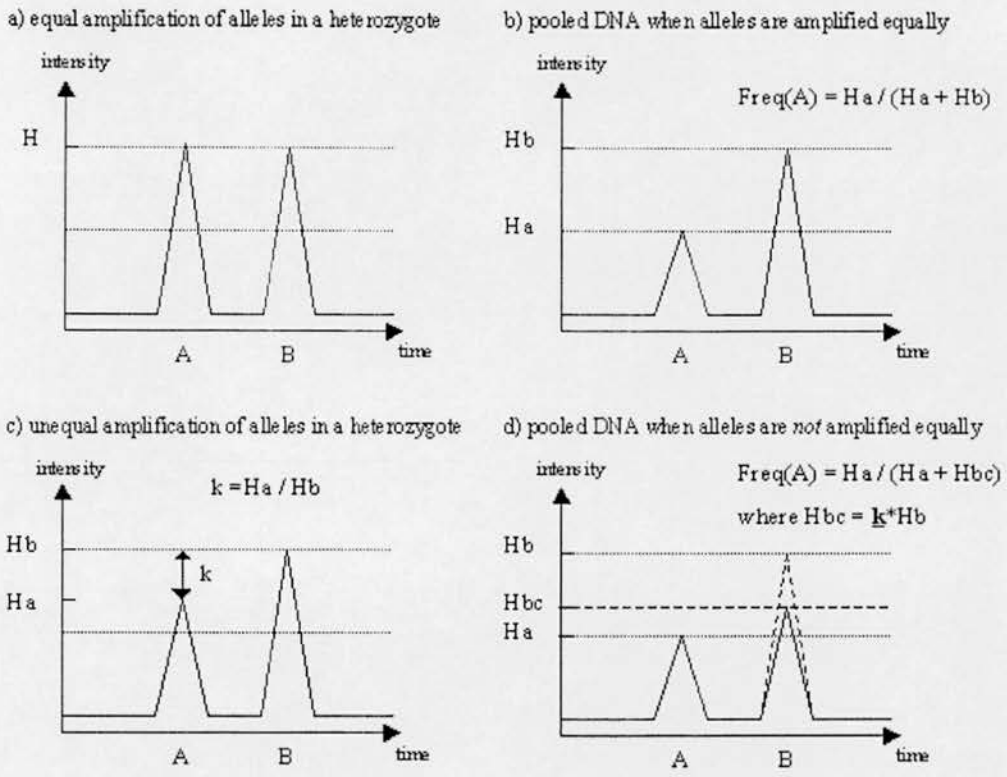


Figure 3.1: Determination of allele frequencies at a SNP locus in pooled DNA. Graphs represent peaks obtained by genotyping a heterozygote and pooled DNA at a locus with alleles A and B. In an ideal situation, both alleles of a heterozygote are amplified equally (a) and allele frequencies in pooled DNA can be derived directly (b). In practice, both alleles are not amplified with the same efficiency (c) and peak heights obtained by genotyping pooled DNA must be corrected prior to deriving allele frequencies. In the example shown in (d), amplification of allele B is more efficient than that of allele A. It follows that the peak height observed for allele B must be reduced by a factor proportional to the difference in the efficiency of the amplification of the two alleles to account for this phenomenon. This factor is taken as the ratio of peak heights obtained for heterozygotes: the correction factor k. H: peak height obtained by genotyping a heterozygote when alleles are amplified equally; Ha: peak height obtained for allele A; Hb: peak height obtained for allele B; k: ratio of peak heights obtained for a heterozygote; Hbc: peak height obtained for allele B, corrected for the better amplification of allele B compared to that of allele A; Freq(A): frequency of allele a in pooled DNA.

3.2 Choice and design of the database and interface

3.2.1 Database and interface requirements

The computational tool had a two-fold aim. It aimed to enable: i) the storage of raw and processed association data, using a database, and ii) the submission and analysis of raw association data, using an interface to the database.

Several types of data required storage:

- the individual samples, groups of individual samples with the same phenotype (e.g. bipolar affective disorder, unipolar depression, schizophrenia or control), heterozygote references and pools of DNA,
- the combination of information on several samples of individual DNAs, or DNA pools,
- the markers tested and protocols for genotyping assays,
- the results of the genotyping assays of individual samples and DNA pools,
- allele frequencies in DNA pools, which are derived from peak heights, and in sets of individual DNAs,
- the tests for allelic association of the alleles at the loci tested with the phenotypes under scrutiny.

The interface to the database aimed to:

- compute allele frequencies in sets of individual DNAs and in DNA pools,
- perform tests of allelic association based on individual genotypes or DNA pools,
- integrate these data to the annotated sequence of the candidate region,
- be user-friendly and hence accessible to any bench scientist involved in the project,
- reduce the need for direct access to the data by users in order to prevent erroneous manipulation of data,

- be flexible to allow: i) decisions by the user to validate the data submitted and analysed, ii) re-submission of data necessary to cope with technical difficulties inherent to large-scale genotyping, and iii) addition of protocols for genotyping assays not yet developed.

The approach developed to solve the problem caused by the wealth of data generated by association studies consisted of modifying the ACeDB database (http://www.acedb.org) to store the information generated, and of creating a CGI front-end to submit, analyse and format these data.

3.2.2 Database choice and design

3.2.2.1 Database choice

The ACeDB database engine is an object-oriented system capable of storing, retrieving and displaying complex biological information (Walsh et al., 1998; Kelley, 2000). The term 'ACeDB' stands for '**A** ***C****aenorhabditis* ***e****legans* **D**ata**B**ase' and reflects the original role of ACeDB in the *C. elegans* sequencing project, namely the coordination of the sequencing effort and the integration of the worm sequence with the genetic and physical maps (Waterston and Sulston, 1995). ACeDB is now used in WormBase, an extension of the *C. elegans* sequencing project (Stein et al., 2001; Chen et al., 2005). Several technologies allow direct access to biological databases. Relational databases such as SyBase (Kirkwood, 2000), Oracle (Mishra and Beaulieu, 2002), and MySQL (Dubois, 1999) enable software programs to consult remote databases using structured query language (SQL) and retrieve the results. These application programming interfaces (APIs) can usually be used for databases from different commercial sources and are operating system independent. However, in such relational databases, a DNA sequence and the data associated with it are scattered among many tables for the sake of efficiency, rendering the design of the database large and difficult (Stein and Thierry-Mieg, 1998). A standard to share data objects across the Internet provided by the Common Object Request Broker Architecture (CORBA) solves this problem (Orfali and Harkey, 1998). However, standardised biological data types for CORBA have been difficult to define and the

development of interfaces based on CORBA is complicated (Stein and Thierry-Mieg, 1998). In contrast, ACEDB does not fragment the biological object and provides simple queries searches, and allows complex queries using the powerful retrieval query language ACEDB Query Language (AQL), similar to SQL (http://www.acedb.org/Software/whelp/AQL/).

Storage of sequences, gene structure, genome features and corresponding annotations has recently become available by using Ensembl (Hubbard et al., 2002) and the University of California at Santa Cruz (UCSC) genome browser (Karolchik et al., 2003). However these tools were not readily available at the time when this project started. In addition, these tools do not provide storage for textual data necessary for manual annotation (Searle et al., 2004).

On the other hand, ACeDB provides intuitive object oriented modelling of biological data designed specifically for handling scientific data flexibly, and a graphical user interface with many displays and tools for genomic data such as a genetic map viewer and a sequence annotation display (Walsh et al., 1998; Kelley, 2000). ACeDB has been adapted by multiple groups for use with various organisms, including the human sequencing projects at the Sanger Centre and the Washington University Genome Sequencing Center (Waterston and Sulston, 1998), the budding yeast genome project, SGD (Cherry et al., 1998), the nematode *Pristionchus pacificus* (Srinivasan et al., 2004), several crop plant and livestock genome analysis projects (Wixon, 2000), and other organisms such as the rice blast fungus *Magnaporthe grisea* (Martin et al., 2002), bacteria (Bergh and Cole, 1994), and trypanosomes (Degrave et al., 1997; Blackwell and Melville, 1999). Several other applications based on ACeDB have been developed for various biological studies, such as sequence motifs in the *Helicobacter pylori* genome (Saunders et al., 1998), neurotoxins (Lebeda, 2004), receptors (Nakata et al., 1999), ATP-binding cassette (ABC) protein domains (Quentin and Fichant, 2000), specific enzymes (Cousin et al., 1998), cell signalling (Igarashi and Kaminuma, 1997; Takai-Igarashi et al., 1998), gene expression in human (Miller et al., 1997) and in *Xenopus laevis* embryos (Pollet et al., 2000), and paralogous regions in Vertebrate genomes (Leveugle et al., 2003). A tool has recently been released for the analysis of complex diseases (Burren et al., 2004).

Last but not least, our laboratory already used an ACeDB database when the work presented here started in early 2001 for two purposes: i) to manage the sequence of the disease susceptibility region on chromosome 4p as BAC clone sequences were being released by the Human Genome Project annotated, and ii) to annotate these sequences (Evans et al., 2001). Because it was required that the allelic association data be stored and integrated with the sequence data and because the latter was already managed using the ACeDB database specifically designed for this purpose, it was decided that the tool to develop for the storage of allelic association data would be build on ACeDB.

3.2.2.2 ACeDB

In ACeDB, a 'model' defines the format of the information content of a 'class'. A class is a group of items (or objects) characterised by the same set of features (Figure 3.2). A model is represented as a tree which harbours five type of nodes: i) tags, which are words that must be enumerated in a separate file, ii) basic data types (integer, float, text), iii) modifiers, to indicate uniqueness of fields by UNIQUE or REPEAT, and to create links between objects, using XREF, iv) class names, which contain a '?' as the first character, such as ?Allele, and v) constructed types names, such as #MACHINE. Tags can also be leaves in the tree.

An instance of a class, that is an object belonging to that class, matches the model and is also represented as a tree, containing tags, data items and pointers to other objects (Figure 3.2). In this object, tags appear in positions set in the model, data types are conform to those in the model and pointers refer to objects of the class indicated in the model. An object does not need to match the entire tree specified by its model, but instead is allowed to match only a subtree in the model. The modifier XREF is used for automatic cross-referencing between objects. It is followed by the name of the tag pointed to. Because XREF is not reciprocal, links between objects can be versatile and two-way double links are specified by XREF in the model of the two classes linked together. Constructed types allow a recursive specification of data within a model. Data fields can be of normal data types such as free text, integers and real numbers, dates, or pointers to other objects in the database.

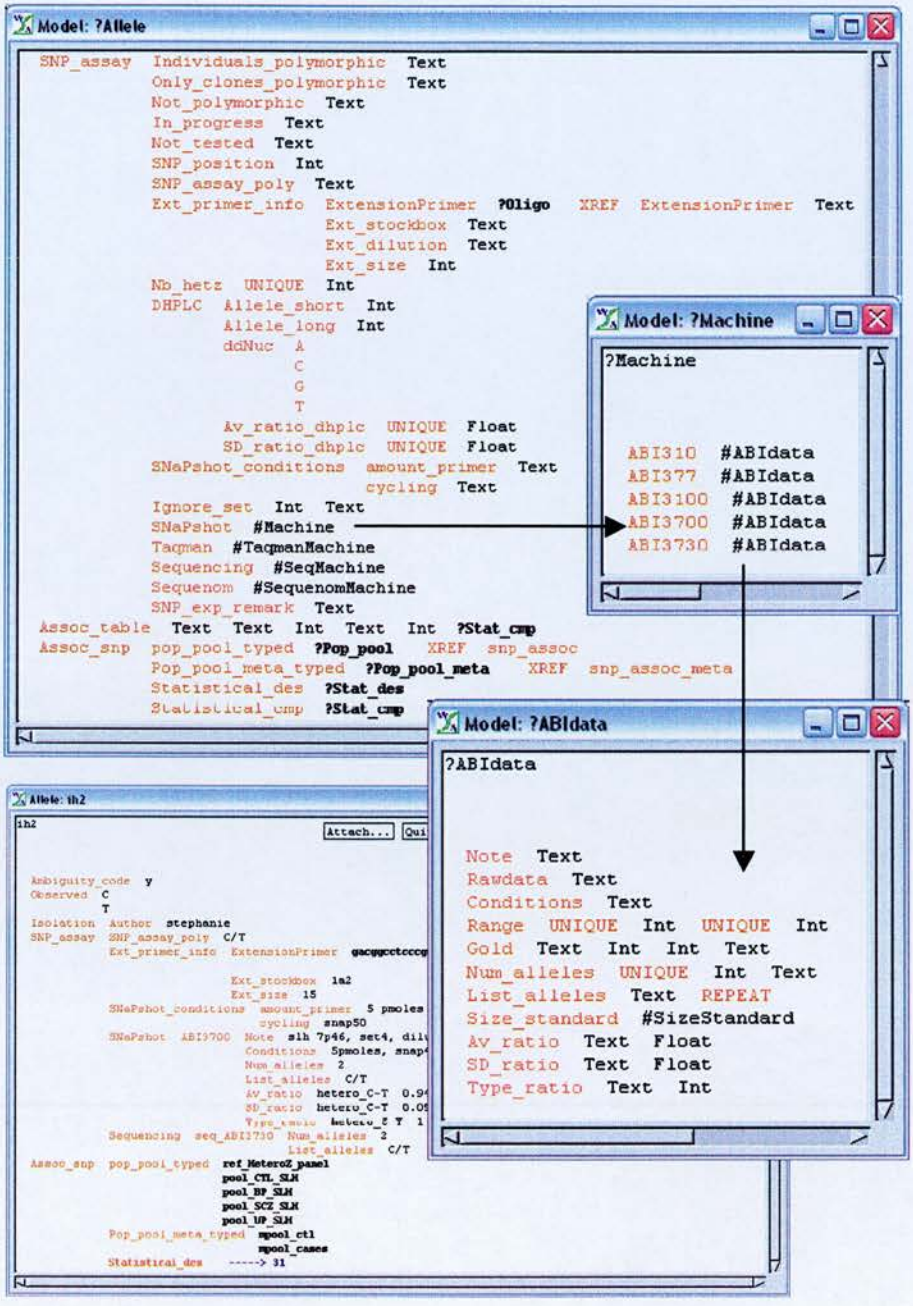


Figure 3.2 : Data representation in ACeDB. The top window shows the model '?Allele' used to store data related to SNP loci, organised as a tree with nodes in brown and leaves in black. The smaller window on the left shows the model '?Machine' used store the name of the genotyping machine used in genotyping assays (tag 'SNaPshot'), themselves stored using in the '?ABIdata' model shown in the bottom right window. The window at the bottom shows an instance of the '?Allele' object, 'ih2'. In this object, tags appear in positions set in the model, data types are conform to those in the model and pointers refer to objects of the class indicated in the model.

ACeDB also provides powerful search tools, such as complex queries using AQL. In particular, the tace program allows text-based access to the ACeDB database. This program is a key component of the tool developed here because it allows queries to be incorporated into scripts for automated consultation of the database.

3.2.2.3 Design of the ACeDB extension for allelic association data

The allelic association data to be stored related to various components of association studies, such as markers and experimental conditions of the genotyping assays, or samples used, as well as raw genotyping results at the level of the individual samples and groups of samples sharing a common feature, such as disease status, and finally, results of tests for association with disease. The design of the database followed naturally from the different types of data to store.

To comply with the settings already existing in the ACeDB database in use to store information of markers and experimental conditions of genotyping assays, it was decided that microsatellite data would be stored in 'STS' objects, while SNP data would be described by 'Allele' objects (Figure 3.2). For both types of locus, these objects would also contain information relating to genotyping assays for individual samples and pooled DNA. However, because the type of data on genotyping assays are similar for both types of marker, and for the different genotyping machines, the data would be stored in a novel constructed type common to both types of markers, such as '#Machine' and '#ABIdata' but accessed from models for both types of markers, 'STS' and 'Allele' (Figure 3.2).

As for samples, three novel models would be used: i) '?Individual', for individuals, ii) '?Pop_pool', for groups of individual samples and for pooled DNA, and iii) '?Pop_pool_meta', for groups of population samples or groups of DNA pools. Individual samples are distinguished from other types of sample because they are entities different from a 'population' or a pool. On the other hand, a population sample and a pool sample would be described in the same manner because they both comprise individuals and their description would for example include the list of individual samples they contain, although they are distinguished in a population sample, but not in a DNA pool. This would be indicated in an '?Individual' object by a pointer after the tag 'Gridded', to a '?Grid' object, which shows the position of

individual samples in 96-well plates for an individual sample used in a population sample, but not for an individual in a pool of DNA. Similarly, a 'meta' sample made of several population sub samples can be described with the same terms as a 'meta' sample made of several pooled DNA sub samples, so both 'meta' would be stored in a 'Pop-pool-meta' object.

Genotypes of individual samples would be stored in the ?Individual model, while a novel model '?Stat_des', for 'statistical description' would store raw genotyping results for both population and DNA pool sample types, as well as results from their analysis, including allele and genotype frequencies. To store details of the analysis, such as a list of peak heights for a given allele obtained for several replicates of a DNA pool genotyping assay, or a list of derived frequencies used to compute the single value of the allele frequency average, a novel model '?List_values' was required. It would store the raw or derived data while excluding them from the display in the '?Stat_des' objects containing the main results of the analysis. It would be accessed from the '?Stat_des' object using links to objects of the '?List_values' class. Finally, results from tests of association of alleles with phenotype would be stored in a novel model, 'Stat-cmp', for 'Statistical comparison', indicating which samples are compared at which locus, based on which set of data.

To allow for easy navigation from an object to another, these models would also provide several links to one another (Figure 3.3). For example, the user would be able to consult results from an association test, (Figure 3.3,'Stat_cmp', bottom right) based on two sets of population samples, and use links in the object to navigate to one of the data sets to consult allele frequencies (Figure 3.3, 'Stat_des' far right), and continue by following links to the individual samples in the set and obtain their genotype (Figure 3.3, 'Individual', top left). Each model is detailed below.

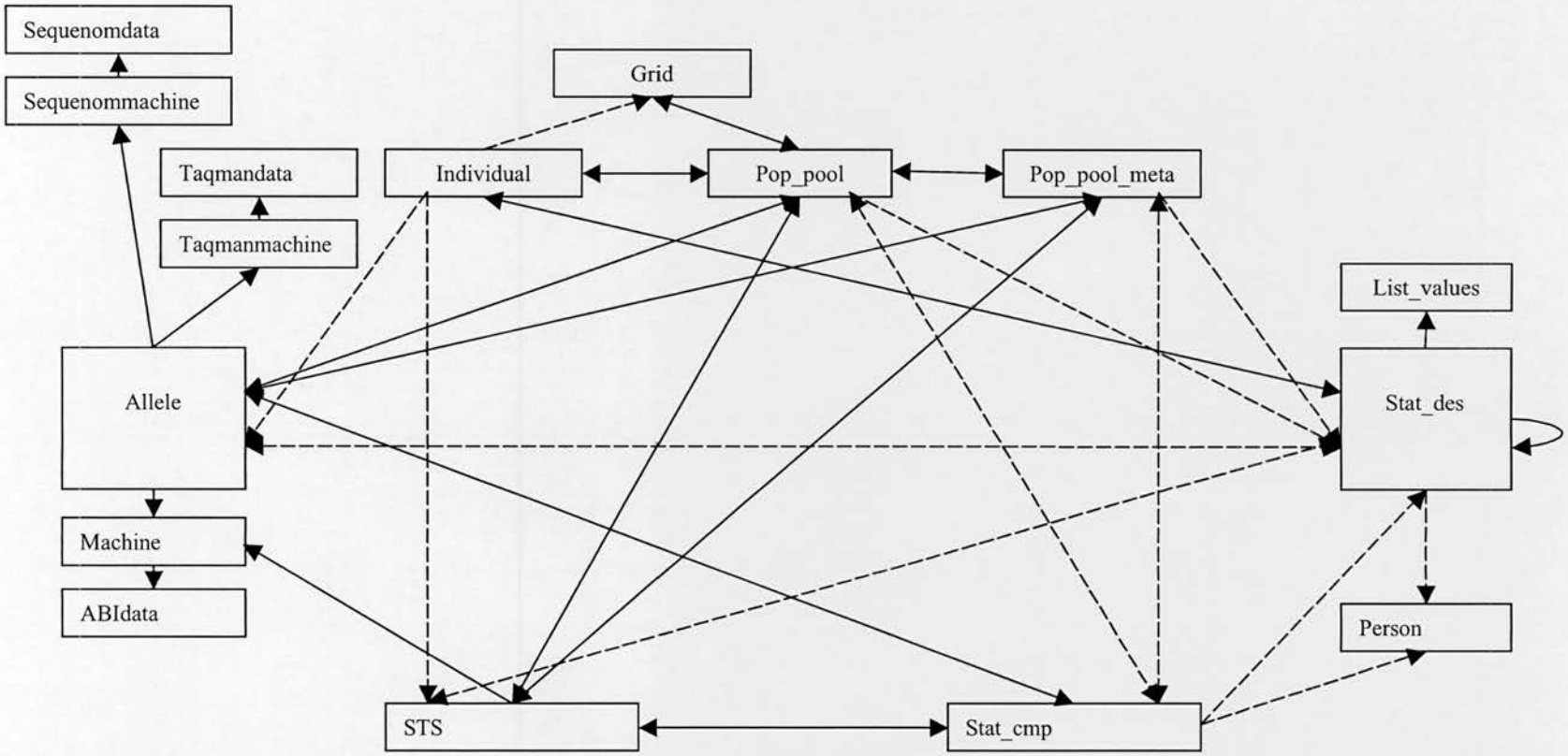


Figure 3.3: Relationships between ACeDB models for allelic association data management. A box represents a model. A solid arrow shows a cross link from one model to the other while a solid double-headed arrow represents a reciprocal cross link between tags of two models. A dashed arrow represents a simple link from one model to another, while a dashed double-headed arrow signifies that a tag from one model has a simple link to another model and vice-versa.

3.2.3 Interface choice and design

To facilitate the analysis and the integration of data into the database, an interface was required. It was decided that this front-end would be developed using Perl and the Common Gateway Interface (CGI) for four reasons. First, Perl is an accessible programming language and is widely used in bioinformatics for local data management, conversation of data format and access to databases (Wall et al., 2000). Second, CGI allows the creation of HTML forms and scripting to analyse submitted information. Third, the combination of Perl and CGI had already been used in various ACeDB-based projects, such as the WebAce (http://webace.sanger.ac.uk) and AceBrowser (Stein and Thierry-Mieg, 1998) web interfaces to the ACEDB database, which both provide basic functionality for searching and navigating ACeDB databases. Finally, our laboratory already used this strategy and had developed Perl/CGI scripts to generate forms for the submission of STS data. A Perl/CGI tool was thus selected to keep the set of tools homogenous and because support for this project would be available, in contrast to alternative strategies based on JAVA, or the generation of HTML pages using PHP.

A CGI script would provide an interface to the novel database and generate HTML forms, thus achieving three goals. First, these forms avoided direct entry of data into the database, reducing the risk for erroneous manipulation of existing and novel data. HTML form fields allowed data to be handled and files created in the correct format, before being read into the database. As finding errors incorporated in the database can prove time consuming, it was concluded that integration of novel data should not be automatic but, instead, verified first by the user. Second, CGI scripts also enabled the retrieval of data of interest from the database, so that the user could only choose appropriate data necessary for the description and analysis of the data submitted. For example, it was important for the sake of reproducibility that peak height ratio values for a pool typed at a given marker were obtained with the experimental conditions stored in the database. Third, the major advantage of these CGI scripts was to allow the analysis of the data submitted. This front-end would allow the submission of DNA pool data and of individual genotypes using a format

similar to the tools used by bench scientists. Furthermore, it would also be able to respond to needs expressed by bench scientists after its initial design. For example, this occurred when they gained access to a facility for high-throughput genotyping assays. Results from these assays are generated in the form of an Excel file that contains data for a dozen of sets of individual samples genotyped at a given marker. A routine was developed and integrated into the software that parses these files and analyses each of the sets automatically. In addition, the front-end would be used to merge DNA pools sets of individual samples into meta samples and analyse the merged data sets. Last but not least, the user would be able to perform statistical analyses of allelic association.

The following sections describe the representation of the allelic association data, as well as the CGI front-end and the ACeDB extension developed to manage the data. The CGI front-end and the ACeBD extension were developed on a Unix workstation and on a PC running Windows XP Home.

The interface design involves seven interacting areas, as indicated by blue ellipses with a white box label in Figure 3.4: i) 'first access', when the user starts a session, she/he is given the choice between several task, ii) 'parameter verification', to test the integrity of selected parameters or submitted data, iii) 'manual submission', to display a form for submission of peak heights, or individual genotypes, iv) 'database query', to consult the database, for example to verify submitted parameters and data, v) 'statistical analysis', to treat submitted data and write file in ACEDB format, vi) 'association study', and vii) 'data integration', to read created files in the database, and subsequently update related objects if necessary. Each area consists of several subroutines. Subroutines mediating the different tasks performed by the interface are highlighted in Figure 3.4, such as 'upload_analysis' for uploading of file generated by high-throughput analysis, 'get_grid' to retrieve the display of the plates used for submission of individual genotypes and heterozygote reference peak heights, 'group_analysis' to analyse data from population, pool and meta samples, 'assoc' to perform association studies, 'cgi_ace_read' to read data into the database, or 'update_analysis' to test whether exisiting objects should be updated using data recently submitted.

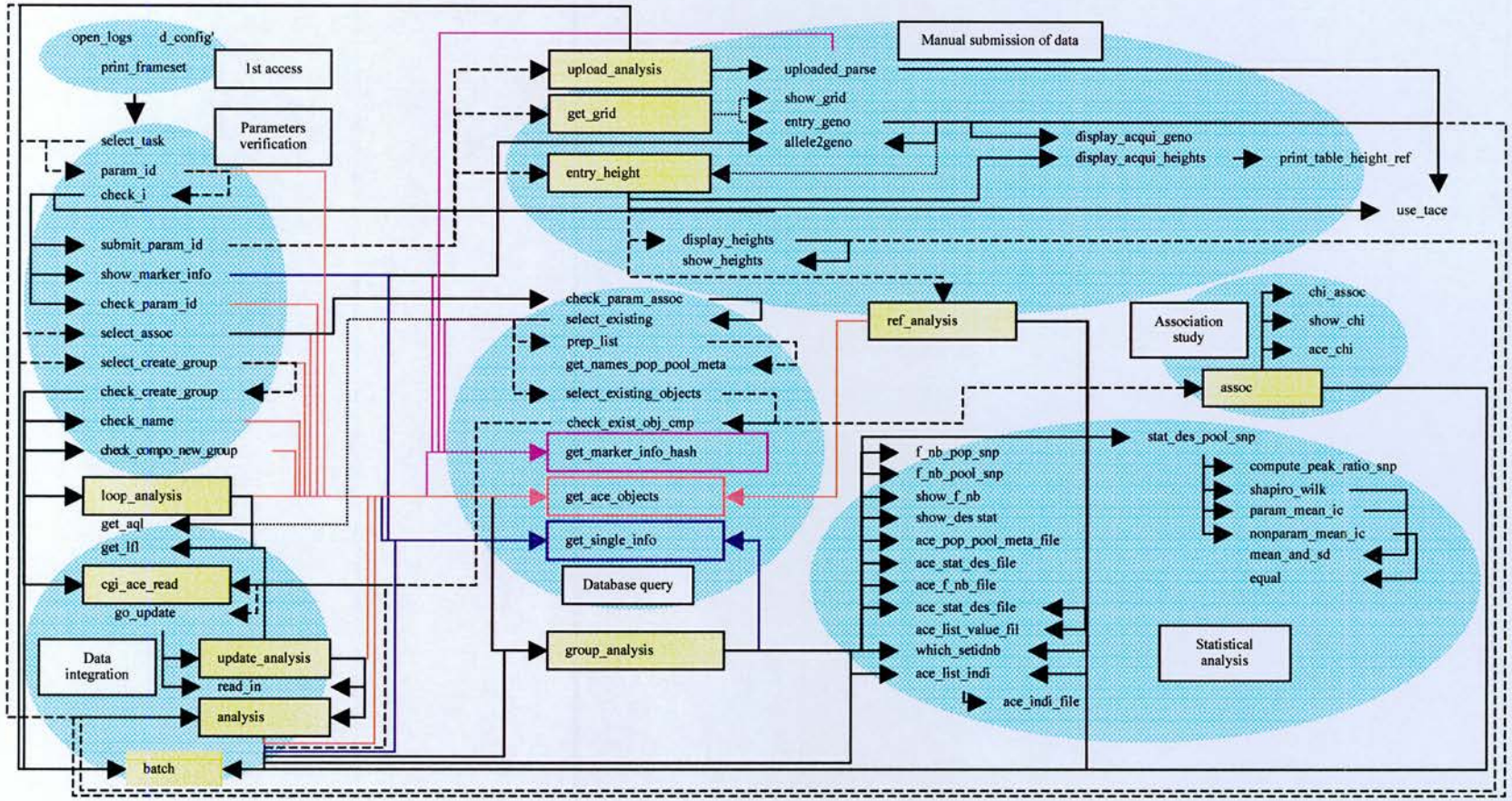


Figure 3.4 : Relationships between subroutines in the script for allelic association data management in ACeDB. A solid arrow represents a direct call of a subroutine by another, and a dashed arrow represents a call via CGI. An ellipse gathers subroutines involved in a same general task indicated in a white text box. Yellow text boxes indicate key subroutines. Boxes with coloured frame and links indicate three subroutines important for database query, and called by many subroutines.

3.3 Protocol for the management of allelic association data

3.3.1 Individual samples

Each individual DNA was represented in the database as a distinct entry. The entry would contain its accession number, disease status and data from studies it was used in.

3.3.2 Sets of individual samples

In the rest of the chapter the term 'population' refers to a group of individual samples with the same phenotype, e.g. disease status. In a population, allele frequencies at any given locus are inferred from the direct counts of alleles in each of the individuals included in that population. Allele frequency differences between populations or groups of populations are analysed using the standard χ^2 test (Sokal and Rohlf, 1997).

3.3.3 Pooled DNAs

In the rest of the chapter the term 'pool' refers to a pool of DNA from individual samples with the same phenotype. Inference of allele frequencies at any given locus in a pool of DNAs requires genotyping a set of heterozygotes at this locus and measurement of the mean ratio of the height of the peak obtained for each of the two alleles (Figure 3.1). Peak heights obtained for each heterozygote, their ratio and results of the analysis of the set of heterozygotes form a data set that is referred to as 'heterozygote reference' hereafter. These data are used to correct for unequal allelic amplification and infer allele frequencies in pools. They must therefore be submitted before peak heights obtained for a DNA pool at a given locus can be analysed. The genotyping assay of a pool of DNAs should be replicated at least four times (Le Hellard et al., 2002). Genotyping of pooled DNA at microsatellite loci is hindered by the number of alleles, and hence peaks to measure, and by slippage of the polymerase, or 'stutter', typical of microsatellite genotyping.

Microsatellite genotyping in pooled DNA is not considered here. The statistical analysis of data obtained for a DNA pool at a locus includes a set of descriptive statistics based on the series of replicate values. The mean peak height ratio and allele frequencies obtained for a pool are characterised by the following statistics: arithmetic mean, median, standard deviation (SD) of the mean, standard error for the mean (SEM), coefficient of variation of the mean, and 95% confidence interval (CI) for the mean or the median. To decide whether the CI should be calculated for the mean or the median, a test for the normal distribution of the allele peak height ratios must be performed. Here, the Shapiro-Wilk test for goodness-of-fit to normal distribution is used (Shapiro and Wilk, 1965). The test has been shown to be a reliable test for normality (Pearson et al., 1977) and can be used for samples with just 3 and up to 2000 individuals (Royston, 1982). If a normal distribution can be assumed, the computation of CI then uses the Student statistic. Otherwise, the non-parametric Wilcoxon test for the symmetry of the dispersion is performed to decide which of the Sign (asymmetry) or Wilcoxon (symmetry) statistics should be used to calculate the CI (Sokal and Rohlf, 1997). Allele frequency differences between pools or samples of pools are analysed using a modified χ^2 test that accounts for the variance of the correction factor computed for the heterozygote reference, and for the variance of allele frequencies derived from replicating the genotyping of pooled DNAs (Le Hellard et al., 2002).

3.3.4 Meta samples

The extension of the ACeDB database and the CGI front-end developed for the management of allelic association data also allowed the meta-analysis of data from several populations, or several DNA pools. Such groups of populations or DNA pools are referred to as 'meta samples' hereafter.

3.4 Submission and analysis of allelic association data

The type of task, e.g. submission of pool data, association study, creation of meta sample, is chosen using a form shown in figure 3.5. Each task is described in the following sections.

3.4.1 Submission of genotyping data

The submission of the new genotypes for a population, a series of peak heights for a heterozygote reference or a pool obtained with a given marker uses four distinct steps: i) selection of parameters to identify the data set to create, ii) submission of the data to be analysed and stored, iii) analysis of the data and display of results, and iv) integration of these data and results of analysis into the database.

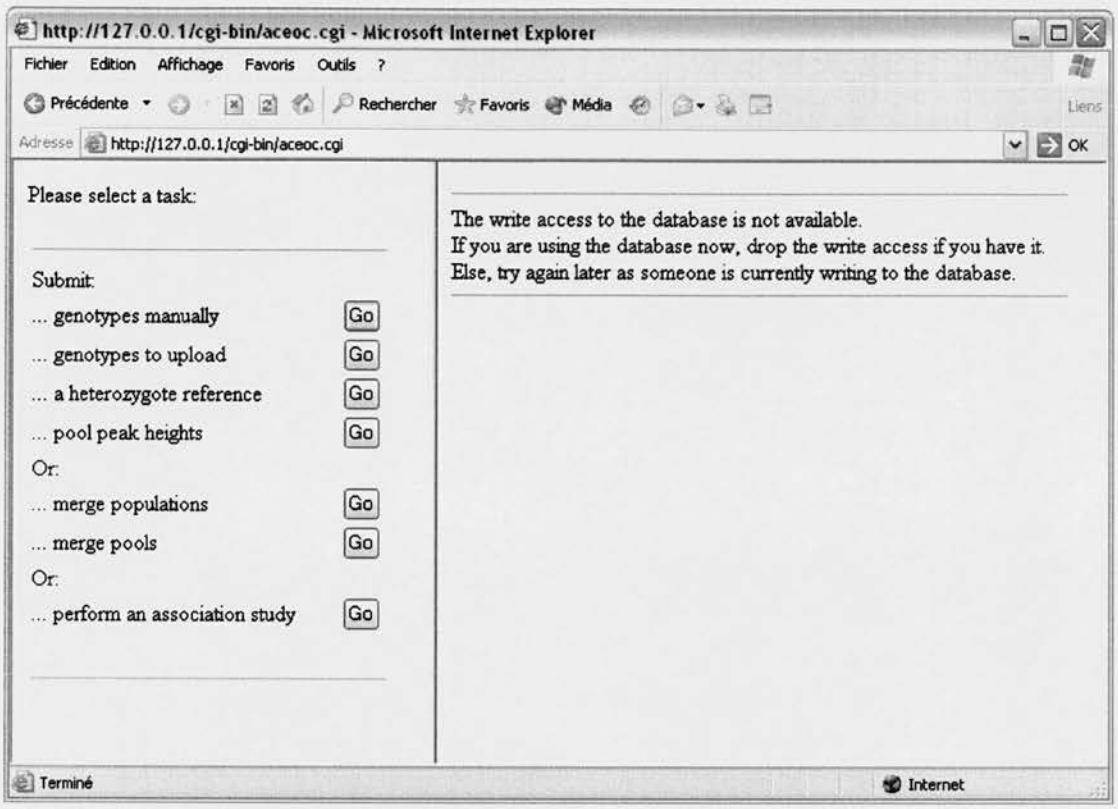


Figure 3.5: CGI form for the selection of task. The first page displayed by the CGI front-end to the ACeDB database contains two frames. The left frame is used to select the task to perform, and, later, parameters necessary for the analysis selected. The right frame is used to display error messages and results of the analysis. For example, the write access to an ACeDB database is granted to only one database user a time. The front-end therefore starts by determining whether it is available. If it is not, the user is warned, as he or she would not be able to read into the database the results of the analysis performed.

3.4.1.1 Identification of data sets

First a CGI form is displayed which requests the submission of the following compulsory parameters: the names of the user, of the population or pool, of the marker and of the genotyping platform, and the number of replicates. These data must be chosen from pull-down lists. Some lists are retrieved from the database using the tace program, e.g. the name of the populations, heterozygote references, pools and markers. After verification of the validity of the parameters submitted, any missing compulsory data is indicated to the user by an error message. Once parameters are correct, the tace program is used to retrieve the data related to the marker (name, polymorphism, position) and the experimental conditions used to obtain data being submitted (genotyping method, name of the oligonucleotide/primer, size of the alleles, number of replicates). These are then displayed to allow the user to verify that they correspond to the conditions used to obtain the current data set (see Figure 3.15 for an example of data stored for a SNP locus). If all data are correct, the user can then submit them. A CGI form is subsequently displayed, the format of which depends on whether the data type relates to a population, heterozygote reference or pool. Analysis of these different data types produces statistical descriptive values, which are common to all types of data analysed.

3.4.1.2 Statistical description of allele frequencies

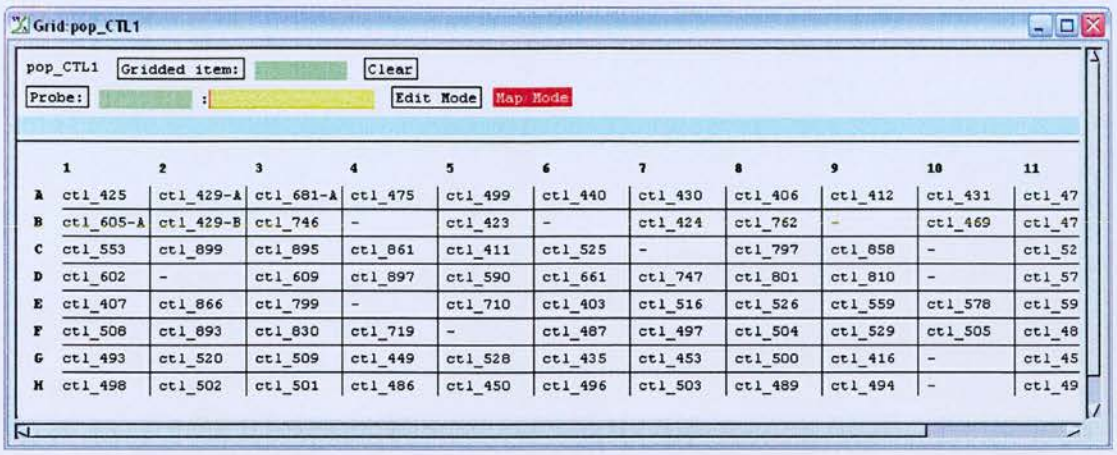
For heterozygote references and pools, the statistical description of submitted peak height ratios or derived allele frequencies consists of the following statistics: arithmetic mean, median, standard deviation (SD) of the mean, standard error for the mean (SEM), coefficient of variation of the mean, and 95% confidence interval (CI) for the mean or the median. The Shapiro-Wilk test for goodness-of-fit to normality of the ratios is performed, and the key values and the decision of the test are displayed. For meta samples of pools, values from each pool are gathered in the single list and analysed as such (section 3.3.1.6).

The allele frequencies in a population, heterozygote reference or pool are then displayed in a table. A file named after the names of the population, reference or pool and the marker is also created in an appropriate '.ace' format so it can to be read into the database. If the user wishes to integrate the data into the database, the file is subsequently read into the database. This causes the creation of a new object in the class named 'Stat_des' ('statistical description'), which contains the values submitted and results of their analysis, such as peak height ratios, allele numbers and frequencies (figure 3.10).

3.4.1.3 Submission of population data

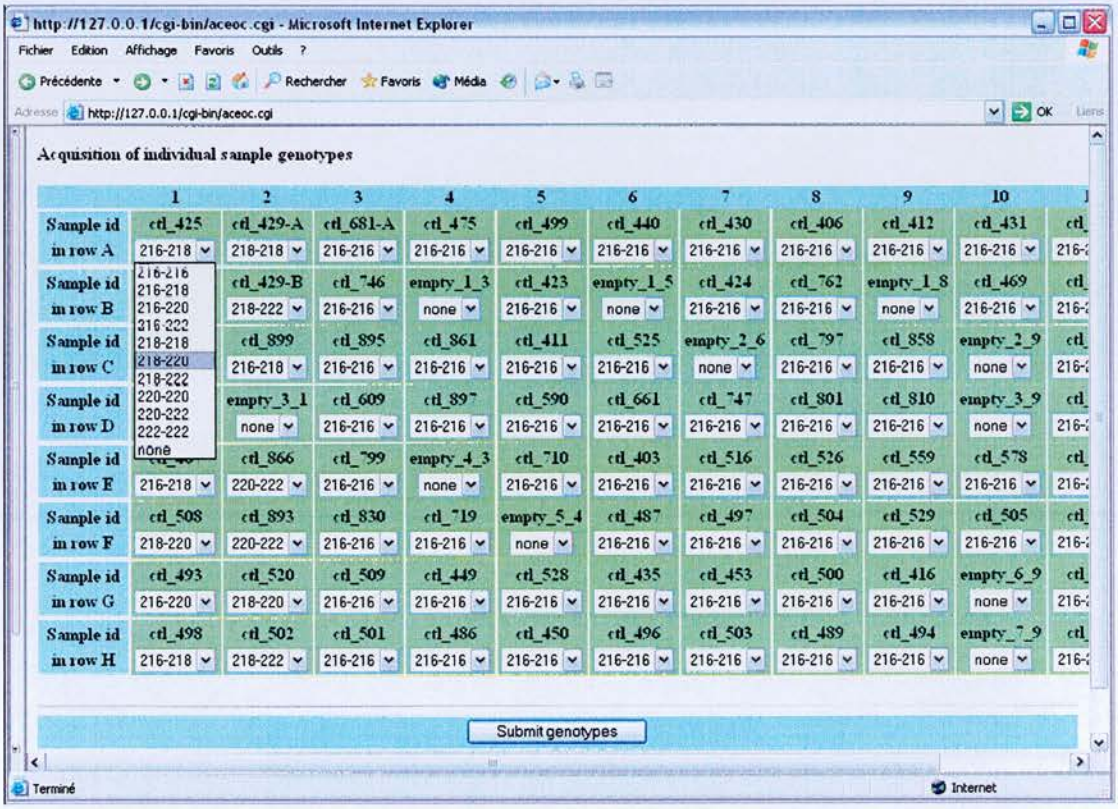
In the laboratory, individual samples are stored in 96-well plates, where a well contains a single sample. In ACeDB, the content of such a plate is stored in a 'Grid' object (Figure 3.6a). In the case of manual submission of individual genotypes, once parameters for the identification of a data set linked to the population are submitted and verified, the CGI script retrieves from the database the accession number and position of samples in the population selected. This information is used to display a form that simulates the format of the 96 well plate. Each well is represented by a pull-down list containing the possible genotypes at the locus of interest (Figure 3.6b). This allows the submission of individual genotypes for use in classical association studies.

a) A population in the ACeDB grid format



	1	2	3	4	5	6	7	8	9	10	11
A	ctl_425	ctl_429-A	ctl_681-A	ctl_475	ctl_499	ctl_440	ctl_430	ctl_406	ctl_412	ctl_431	ctl_47
B	ctl_605-A	ctl_429-B	ctl_746	-	ctl_423	-	ctl_424	ctl_762	-	ctl_469	ctl_47
C	ctl_553	ctl_899	ctl_895	ctl_861	ctl_411	ctl_525	-	ctl_797	ctl_858	-	ctl_52
D	ctl_602	-	ctl_609	ctl_897	ctl_590	ctl_661	ctl_747	ctl_801	ctl_810	-	ctl_57
E	ctl_407	ctl_866	ctl_799	-	ctl_710	ctl_403	ctl_516	ctl_526	ctl_559	ctl_578	ctl_59
F	ctl_508	ctl_893	ctl_830	ctl_719	-	ctl_487	ctl_497	ctl_504	ctl_529	ctl_505	ctl_48
G	ctl_493	ctl_520	ctl_509	ctl_449	ctl_528	ctl_435	ctl_453	ctl_500	ctl_416	-	ctl_45
H	ctl_498	ctl_502	ctl_501	ctl_486	ctl_450	ctl_496	ctl_503	ctl_489	ctl_494	-	ctl_49

b) Form for submission of genotypes for the individuals in the population



	1	2	3	4	5	6	7	8	9	10
Sample id in row A	ctl_425 216-218	ctl_429-A 218-218	ctl_681-A 216-216	ctl_475 216-216	ctl_499 216-216	ctl_440 216-216	ctl_430 216-216	ctl_406 216-216	ctl_412 216-216	ctl_431 216-216
Sample id in row B		ctl_429-B 218-222	ctl_746 216-216	empty_1_3 none	ctl_423 216-216	empty_1_5 none	ctl_424 216-216	ctl_762 216-216	empty_1_8 none	ctl_469 216-216
Sample id in row C		ctl_899 216-218	ctl_895 216-216	ctl_861 216-216	ctl_411 216-216	ctl_525 216-216	empty_2_6 none	ctl_797 216-216	ctl_858 216-216	empty_2_9 none
Sample id in row D		empty_3_1 none	ctl_609 216-216	ctl_897 216-216	ctl_590 216-216	ctl_661 216-216	ctl_747 216-216	ctl_801 216-216	ctl_810 216-216	empty_3_9 none
Sample id in row E		ctl_866 220-222	ctl_799 216-216	empty_4_3 none	ctl_710 216-216	ctl_403 216-216	ctl_516 216-216	ctl_526 216-216	ctl_559 216-216	ctl_578 216-216
Sample id in row F	ctl_508 218-220	ctl_893 220-222	ctl_830 216-216	ctl_719 216-216	empty_5_4 none	ctl_487 216-216	ctl_497 216-216	ctl_504 216-216	ctl_529 216-216	ctl_505 216-216
Sample id in row G	ctl_493 216-220	ctl_520 218-220	ctl_509 216-216	ctl_449 216-216	ctl_528 216-216	ctl_435 216-216	ctl_453 216-216	ctl_500 216-216	ctl_416 216-216	empty_6_9 none
Sample id in row H	ctl_498 216-218	ctl_502 218-222	ctl_501 216-216	ctl_486 216-216	ctl_450 216-216	ctl_496 216-216	ctl_503 216-216	ctl_489 216-216	ctl_494 216-216	empty_7_9 none

Figure 3.6: Manual submission of individual genotypes. The form simulates the format of the ACeDB 'Grid' object that stores the location of the individual samples in the plate used for genotyping assays (a). In the form, each individual sample is identified by its name (e.g. 'ctl_425' in row A, column 1) and its genotype is selected from a pull-down list of genotypes possible at the studied locus (b).

To submit genotypes for several populations obtained via high-throughput genotyping and stored in an Excel file, the user first selects the file to upload. This content of the file is then parsed and data for each population are analysed. Results of each analyses are then displayed (Figure 3.7).

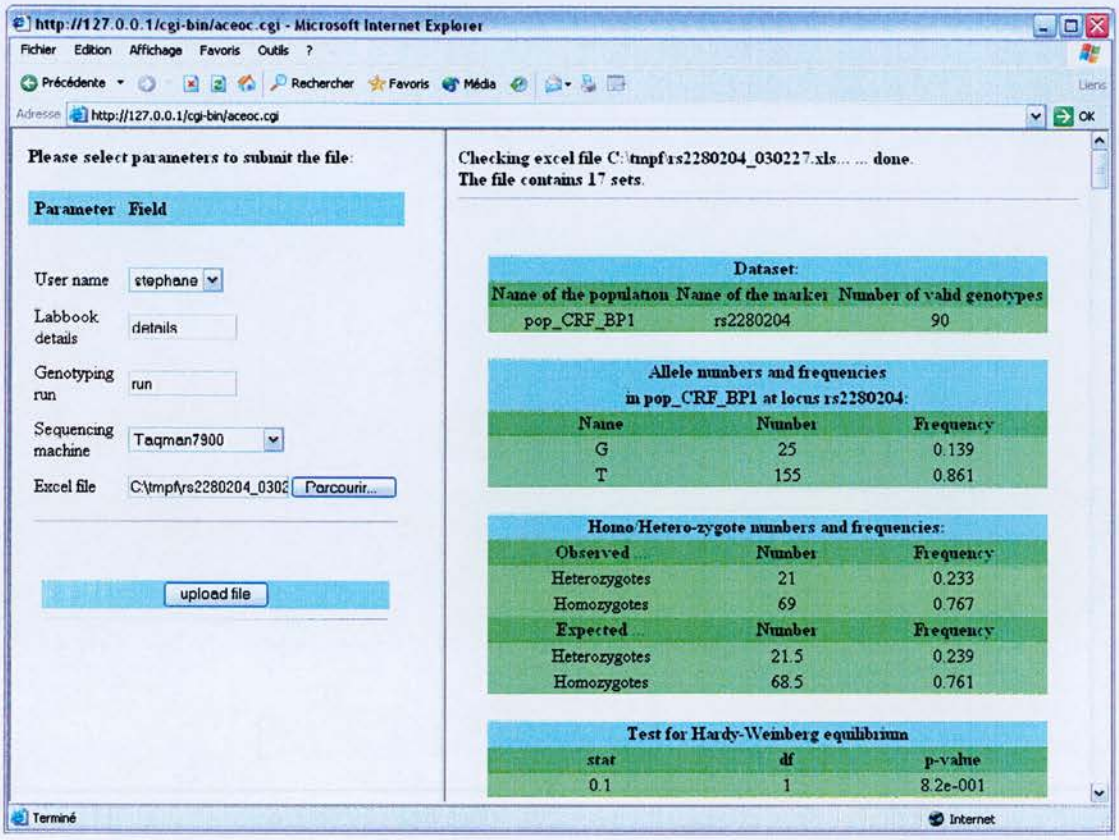


Figure 3.7: Uploading of individual genotypes. The file to upload is selected using the form in the left frame. Results of the analysis are displayed on the right frame for each set contained in the uploaded file. These include the numbers and frequencies of alleles and details of the test for the Hardy-Weinberg equilibrium, which is used to determine whether the data submitted are valid, as shown here for the set of genotypes 'pop_CRF_BP1'.

3.4.1.4 Submission of heterozygote reference data

Heterozygote reference data are obtained by genotyping a set of individual samples and recording peak heights. After selection of the locus and genotyping platform, the form to submit individual genotypes (Figure 3.6b) is used to indicate which of these individuals are heterozygous at the locus studied. A form similar to that for the submission of individual genotypes is then displayed for the user to submit peak heights obtained for these heterozygotes. In this form, the pull-down list used to select the genotype for each individual is replaced by two text fields, one for each peak height to submit (Figure 3.8). The peak heights submitted are then checked. If correct, they are used to compute the correction factor used to analyse data obtained for pooled DNA at the locus studied.

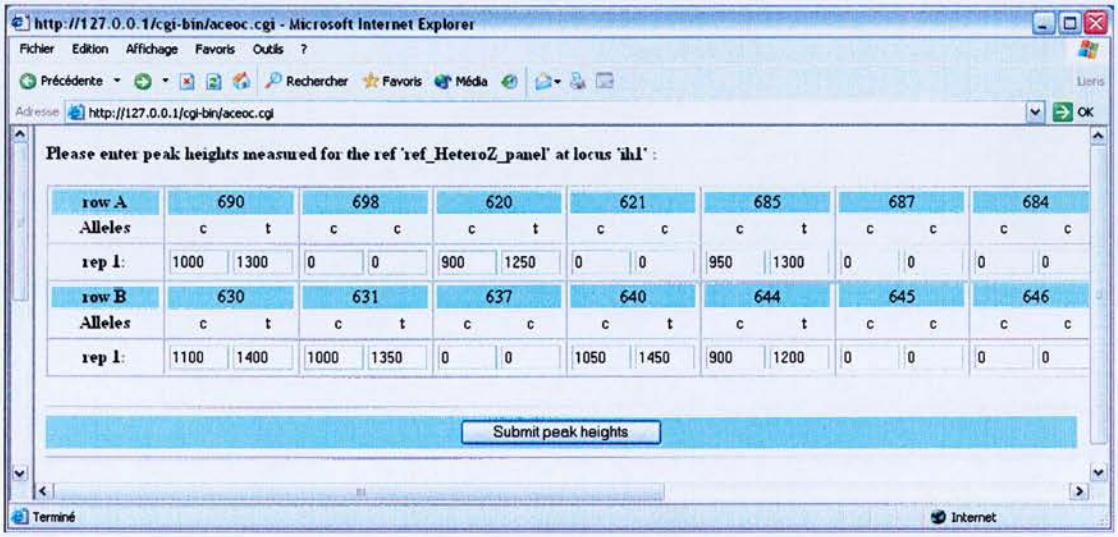


Figure 3.8: Submission of peak heights for heterozygote individuals. The form simulates the format of the plate containing the samples used to detect heterozygotes. The first column indicates the name of the row in the plate (eg 'row A'). Each row in the plate is represented by several rows in the form. The first row indicates the name of the sample (eg '690', '698', ...), while the second row displays the name of the alleles detected (e.g. 'c' and 't'), based on data stored in the database for the studied locus (tag 'ABI3700', sub-tag 'List_alleles' in Figure 3.15). The following rows each correspond to one replicate (e.g. 'rep 1', for 'replicate 1'. Here, the genotyping assay was performed only once). For each individual, a text field is displayed for each of the two alleles, for the user to indicate the peak height observed. Text fields for homozygotes contain the default value '0' as these individuals are not included in the analysis.

3.4.1.5 Submission of pooled DNA data

As for individual genotypes, submission of pooled DNA data first requires the selection of parameters, such as the locus studied, the genotyping platform used and the number of replicates performed. The correct completion of the form is then verified. If no error is detected the database is consulted to retrieve the heterozygote reference for the marker and genotyping platform selected. If no such set can be found, an error message prompts the user to submit data for a heterozygote reference prior to submitting pooled DNA peak heights. If several heterozygote references are found the user selects one of them, via a pull-down list, which is later used to infer allele frequencies in the pool of DNAs. Figure 3.9 shows the form used to submit peak heights. The validity of the values submitted is then verified. Allele frequencies and numbers in the pool of DNAs are finally derived and displayed in a table (Figure 3.10).

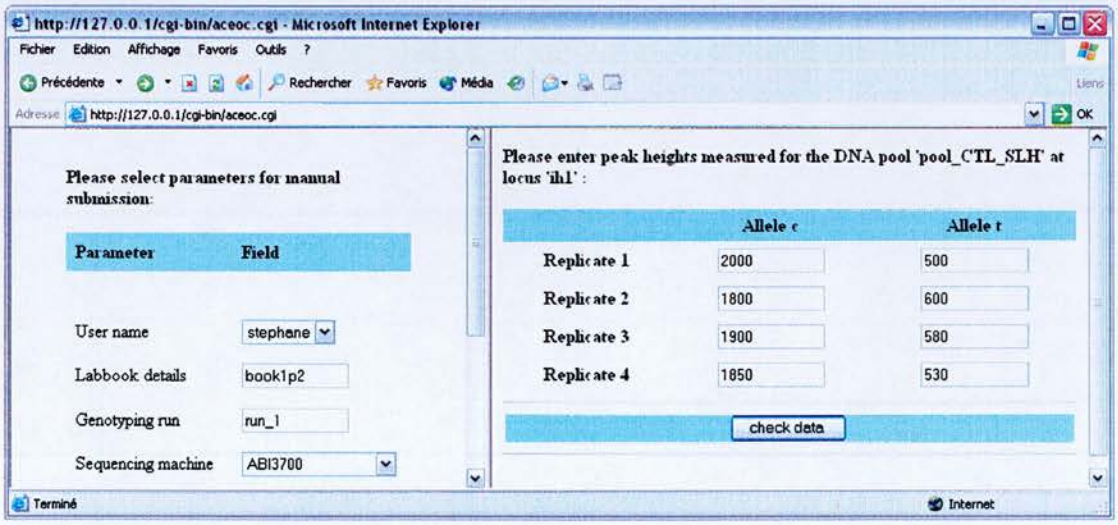


Figure 3.9: Submission of peak heights obtained for pooled DNA. Peak heights obtained by typing pooled DNA (e.g. 'pool_CTL_SLH') at a locus (e.g. 'ih1') using a given genotyping platform (e.g. 'ABI3700') are submitted via the form displayed in the right frame. Each row represents a replicate and contains a text field for each of the allele detected with the assay.

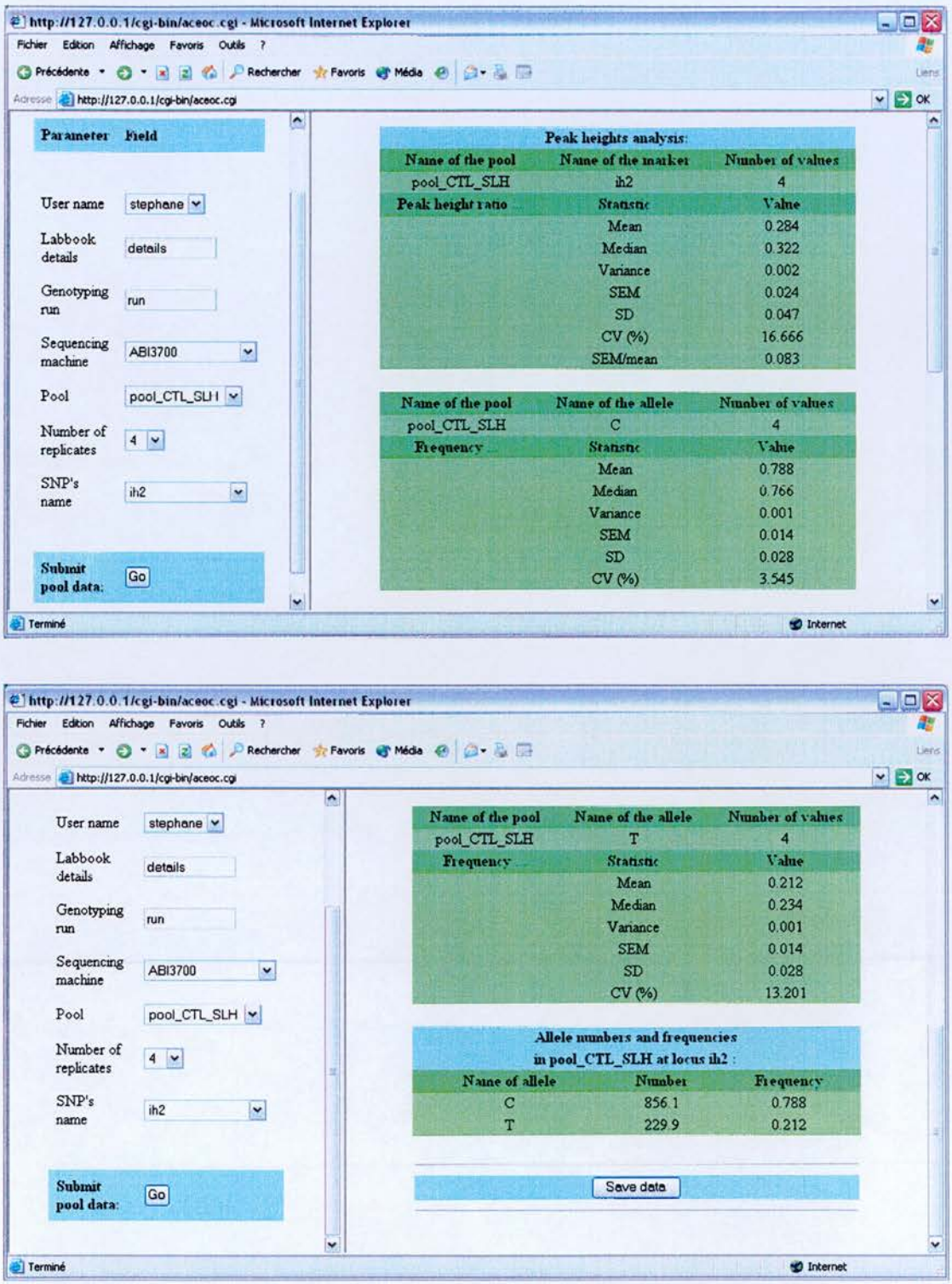
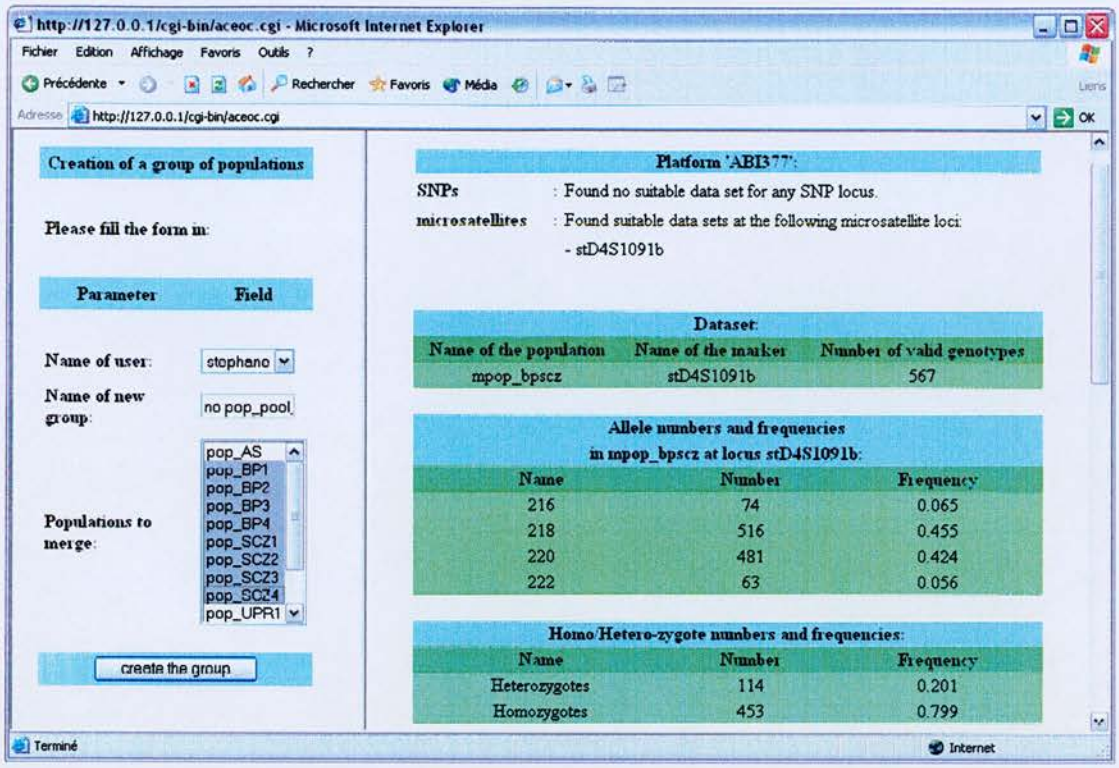


Figure 3.10: Analysis of pooled DNA peak heights. Results of the analysis of peak heights obtained for a pool of DNAs (e.g. 'pool_CTL_SLH') at a locus (e.g. 'ih2') with a given genotyping platform (e.g. 'ABI3700'), based on the selected heterozygote reference, are displayed in four tables. These are shown here in two screen shots. The first three tables display descriptive statistics of the peak height ratio and of the frequency of both alleles. The fourth table indicates the frequency and number of each allele in the pool.

3.4.1.6 Creation of meta samples

Data sets from several populations or DNA pools can be merged and analysed as a single data set, referred to as a 'meta sample'. This is necessary, for example, when an association study has been carried out on several pools of DNA from individuals of one phenotype, which must be combined before comparison with the allele frequencies obtained for individuals with a second phenotype. Single data sets to merge must relate to the same locus and have been obtained with the same genotyping machine. A form is first displayed for the user to indicate the name of the novel meta sample and to select the populations or pools to merge. An error message is displayed if the name is already used, or if a meta sample containing the same group of populations or pools submitted already exists in the database. If this is not the case, the database is consulted to identify data sets suitable for merging. Figure 3.11 illustrates the creation of meta samples. Meta samples are created by indicating the name of the meta sample to create and by selecting populations or pools to merge from a list of populations or pools. The database is then consulted to determine, for each genotyping platform, the list of markers for which data sets exist for each of the populations or pools to merge. If suitable data sets are found, these are merged and analysed as a single set. As shown here for the meta sample of population of patients affected by BPAD (pop_BP1, pop_BP2, ...) or schizophrenia (pop_SCZ1, pop_SCZ2, ...) and for the genotyping platform 'ABI377', no data sets could be found at SNP loci, but a series of data sets were found for a microsatellite locus and analysed as a single set (a). If no data sets can be found for any marker with any platform, the meta sample can still be created, as shown here for a meta sample of pools of DNAs from patient suffering from unipolar depression (pool_UP), BPAD (pool_BP) or schizophrenia (SCZ) (b), and updated as data are later submitted (see section 3.3.1.7). Analysis and integration in the database of these data lead to the creation of a 'pop_pool_meta' object and of objects for the results of their analysis (Figure 3.18 and 3.19).

a) Creation of a 'meta sample' of populations



b) Creation of a 'meta sample' of pools

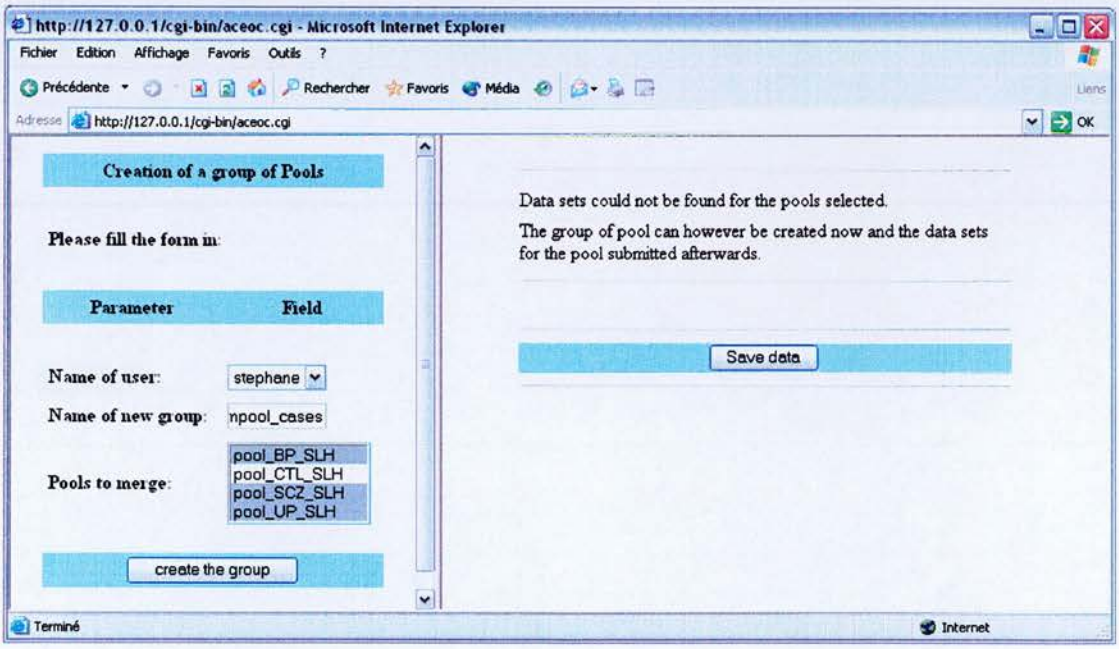


Figure 3.11: Creation of meta samples.

3.4.1.7 Updating meta samples

When individual genotypes or pool data are submitted, analysed and read into the database, the CGI script determines whether the novel data set can be used to update meta samples the population or pool is a member of, if any. For each of these meta samples, the novel data sets are merged with those of the other populations or pools in the meta sample, and the merged data set thereby created is then analysed (Figure 3.12).

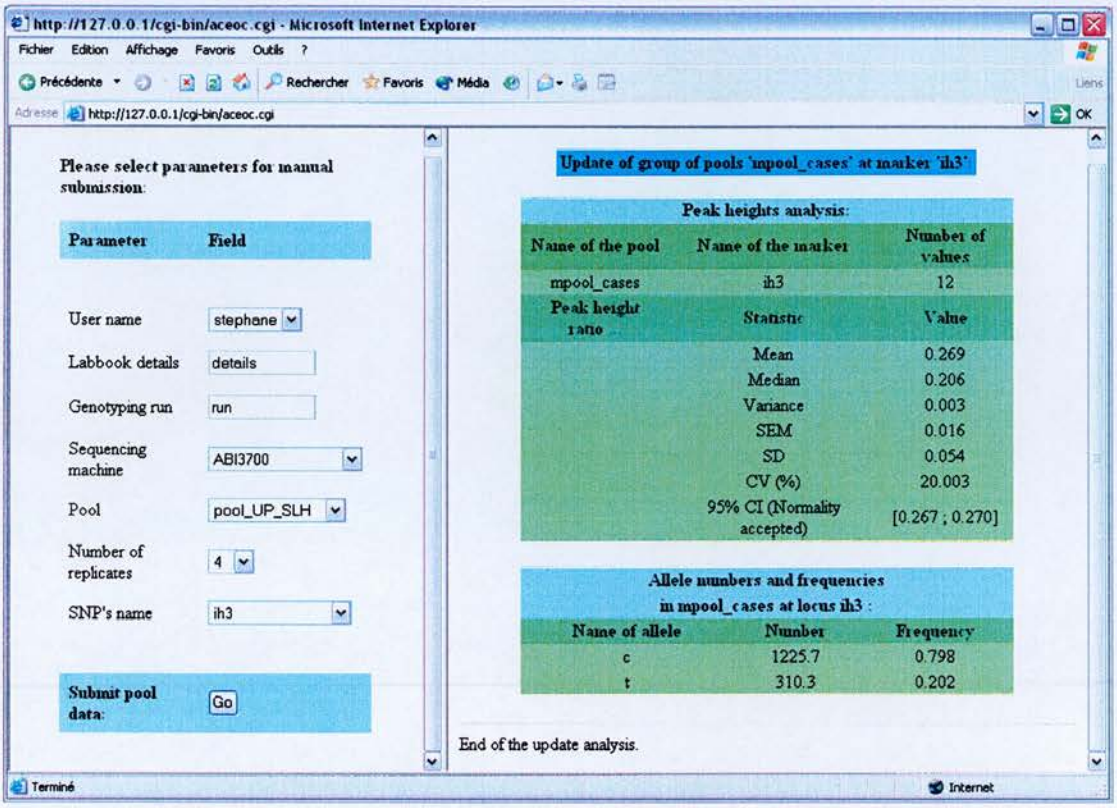


Figure 3.12: Update of meta sample. Peak heights obtained for the pool 'pool_UP' at the SNP locus 'ih3' with the genotyping platform 'ABI377' (see left frame) was submitted, analysed, and read into the database. Prior to submitting these data, a meta sample of pools, 'mpool_cases' was created (Figure 3.11b), and peak heights obtained for pools 'pool_BP' and 'pool_SCZ' for 'ih3' with platform 'ABI377' were submitted, analysed, and read into the database. After integration of the data set for 'pool_UP', the database was automatically consulted to identify the meta samples that 'pool_UP' belongs to, and for each of those, determine data sets that are also available for the other pools in the meta sample. This search led to the identification of a data set obtained for 'ih3' with 'ABI377' for all three pools in the meta sample 'mpool_cases'. These data sets were then merged and analysed as a single set (right frame).

3.4.2 Allelic association studies

The user selects parameters necessary for the identification of the data set, such as his or her name, the SNP assayed, and whether the comparison will be done between populations, pools, or meta samples. The user then chooses the samples to compare from pull-down lists. These lists were derived by consulting the database to determine the samples for which data sets are available for the marker and genotyping platform selected. If the populations, pools or meta samples chosen have already been compared, the user is informed and given the opportunity to abort the analysis or change his or her selection. Otherwise, a modified χ^2 test is performed on the allele frequencies. Results are displayed in a table on the screen (figure 3.13) and an output file is created in the 'ace' format, which is named after the names of each population, pool or meta sample, marker, and genotyping platform. Data are stored in the database if the user so wishes after verification of the results. The file contains the results of the comparison, which will be stored in a database object of the class 'Stat_cmp' ('statistical comparison', figure 3.21). It also contains information that will be stored in each object related to the analysis, e.g. the population, pool or meta sample.

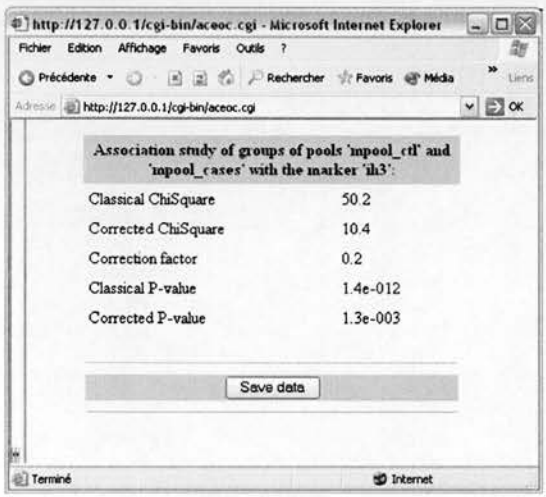


Figure 3.13: Association studies. This figure shows results of the association studies based on meta samples of pools 'mpool_ctl' and 'mpool_cases' at the SNP locus 'ih3': the value of the classical χ^2 statistic obtained without correcting for errors introduced by the use of allele numbers inferred from pooled DNA rather than directly counted by genotyping individual samples, the value of the modified χ^2 statistic accounting for errors introduced by computing allele numbers in pooled DNA, the value of the correction factor used to derive it, and the p-value associated with these two statistics.

3.5 Storage of allelic association studies data

3.5.1 Allelic association data management in ACeDB

Curating allelic association data in the ACeDB database required a adaptation of the architecture of the database and various models were therefore modified, e.g. to allow the management of data relating to the markers typed. Also, new models were created to store data on populations, heterozygote references, pools and meta samples of populations or pools. Novel models also allowed the storage of results of the statistical description of the genotyping data for a given population, heterozygote reference, pool or meta sample with a given marker. Additional novel models allowed storage of results of the comparison of two populations, two pools or two meta samples typed at a marker locus.

3.5.2 Main window

The main window is split horizontally in two sections (Figure 3.14). The upper section is common to ACeDB databases and provides access to tasks available, such as integration of data from file (Edit button) or consultation of the database (Query button). The lower section is novel and provides links to four columns of objects modified or created to manage allelic association data. Clicking on a link brings a window listing the objects of that type, which can be queried in more detail. The list of links displayed is configurable.

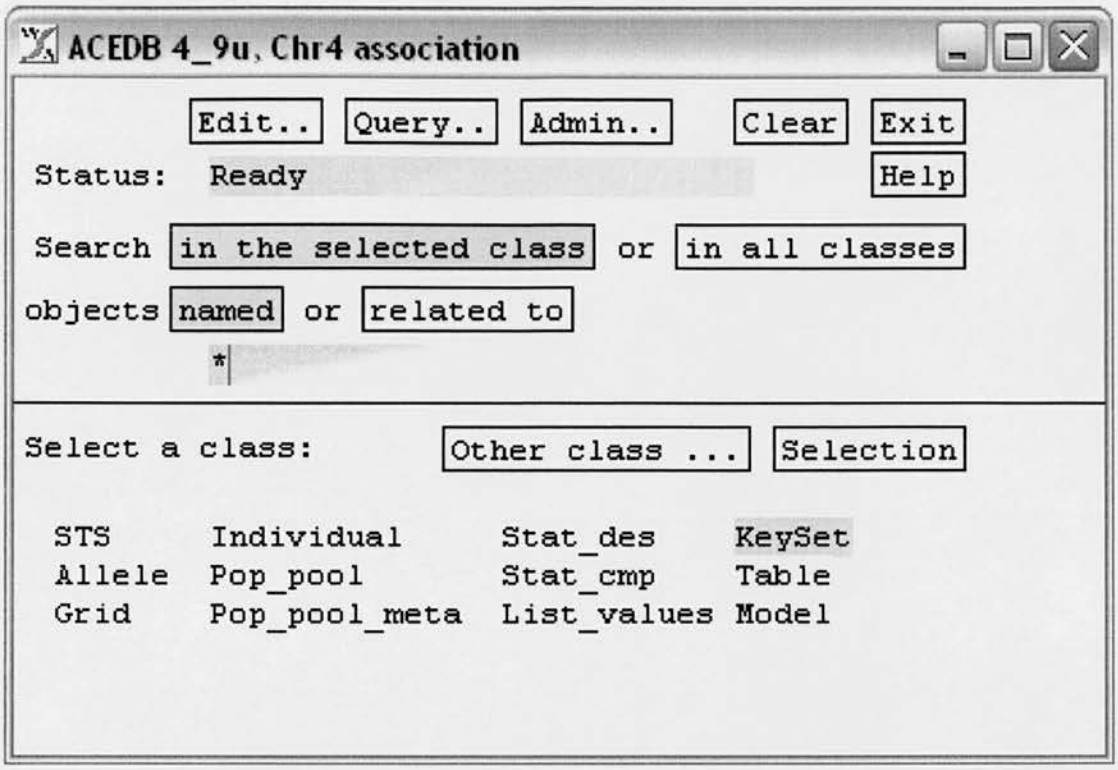


Figure 3.14: Main window of the ACeDB database for allelic association data. The lower section displays links to frequently consulted classes. The left most column has links to markers (STS and Allele) and to grids. Grids provide information on the location of the individual samples on the plates used in genotyping assays. The second column provides access to the samples: i) individual samples (Individual), ii) populations and pools (Pop_pool) and iii) 'meta' datasets that result from the merging of data sets from several populations or pools (Pop_pool_meta). Results of the analysis of genotyping assays of the samples (population, pool and merged set) and results of association studies can be viewed via links in the third column. Saved sets of objects (Keyset), results of a previous query (Table) and the definition of the objects (Models) can be accessed through links in the fourth column.

3.5.3 Modified models

The 'STS' and 'Allele' models were modified to store information on microsatellites and SNPs, respectively (Figure 3.15). With respect to the information managed, the main difference between these STS and Allele classes resides in the storage of experimental conditions. However, data on genotyping assays are stored via a single novel model "Machine" (see below) used for both types of markers. Also, both models provide links to the populations, pools (see model 'Pop_pool' below) and meta samples (see model 'Pop_pool_meta' below) that have been typed at these loci, and the statistical analysis of the results obtained (see models 'List_values', 'Stat_des' and 'Stat_comp' below).

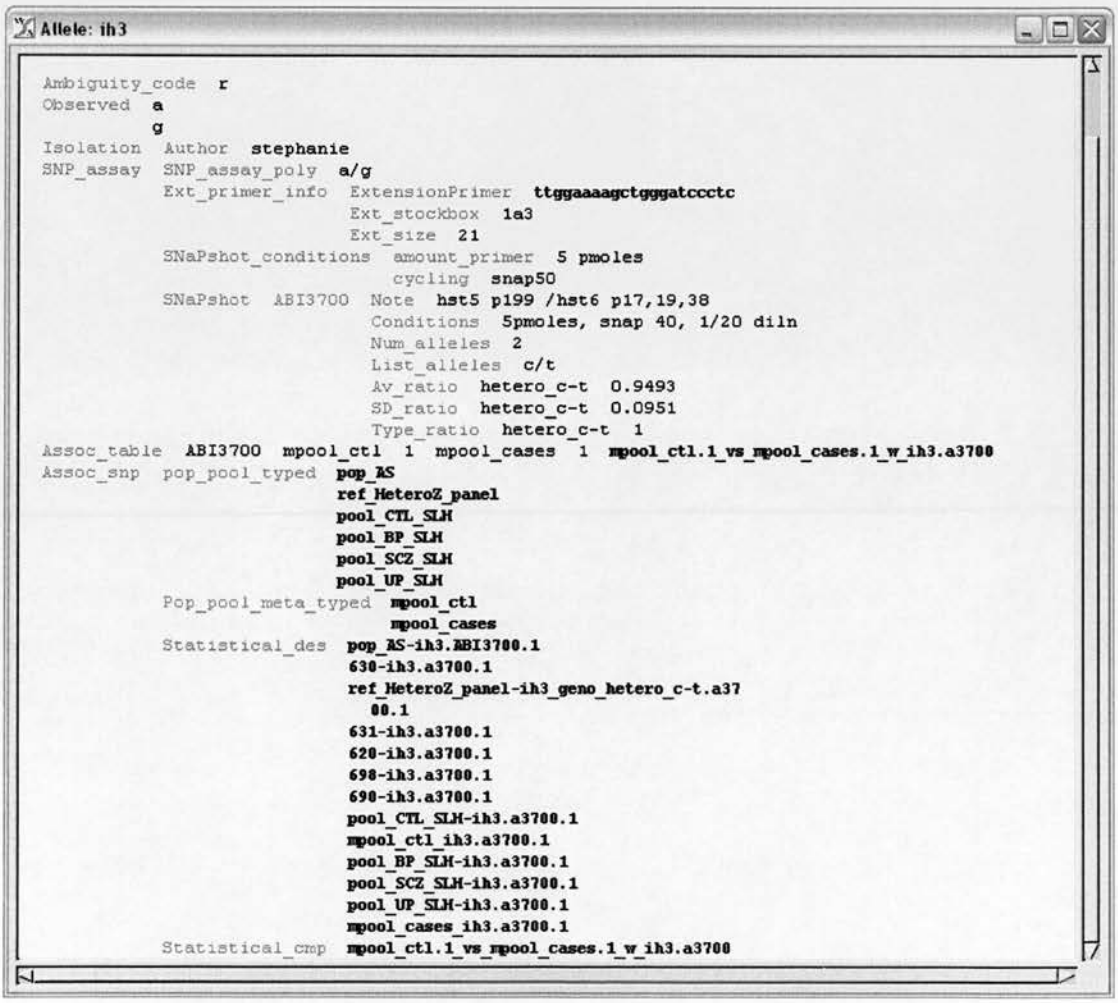


Figure 3.15: Management of SNP data in ACeDB.

Figure 3.15 shows the data stored for the SNP 'ih3'. Data on SNPs are stored in 'Allele' objects. This example shows data stored for 'ih3'. The tag 'SNP_assay' displays data on the genotyping assay of this SNP. In particular, tags 'SNaPshot' and 'ABI3700' concern the assay performed on a 'ABI3700' platform and display the mean peak height ratio obtained for the heterozygote reference (tag 'Av_ratio'). The latter is the correction factor, k, used to account for unequal allelic amplification when inferring allele numbers in pooled DNA. The tag 'assoc_table' displays the list of populations, pools or meta samples included in association studies for that locus. Each entry comprises six columns: i) the name of the platform used to obtain allele numbers (e.g. ABI3700), ii) the name of the first sample (e.g. 'mpool_ctl'), iii) the accession number of the data set used in this analysis (e.g. '1', or '2', '3', and so on as more recent data sets are added to the database for that sample, marker and platform), iv) the name of the second sample (e.g. 'mpool_cases'), v) the accession number of the data set for this second sample (e.g. '1'), and vi) the link to the 'Stat_cmp' object that stores results of the association study (mpool_ctl.1_vs_mpool_cases.1_w_ih3.a3700). The tag 'Assoc_snp' and sub-tags to its right provide links to objects related to the marker 'ih3' such as the object storing data on the samples, tags 'pop_pool_typed' and 'pop_pool_meta_typed', or results of their analysis at that locus, tag 'Statistical_des'.

3.5.4 Novel models

3.5.4.1 Genotyping assays

The novel model 'Machine' stores the list of genotyping platforms used in the laboratory. It displays the experimental conditions used for genotyping samples using these platforms. This data is stored by the novel model 'ABIdata'. An example of such data for the genotyping of the SNP 'ih3' on the ABI3700 sequencing machine is shown in figure 3.15 to the left of the tag 'SNaPshot': tag 'ABI3700'.

3.5.4.2 Individual samples

The 'Individual' model was created to record which population, pool and meta sample an individual DNA sample belongs to, with links to these objects and the Grid object for populations in which the sample is included. This model also indicates the genotype of the sample at SNPs and microsatellite loci, and provides links to these markers (Figure 3.16).

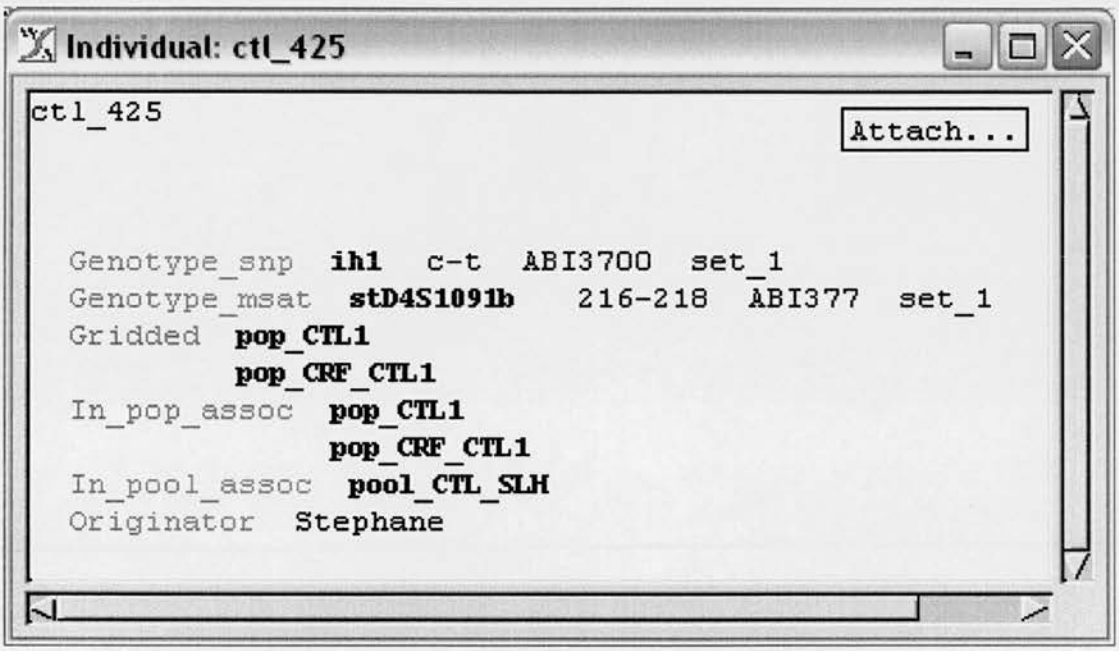


Figure 3.16: DNA sample data management in ACeDB. Example of the sample 'ctl_425' indicating that this individual is heterozygote c-t at SNP 'ih3' and 216-218 at microsatellite 'stD4S1091b', and included in the population of control samples 'pop_CTL1' and that its DNA was used in the pool 'pool_CTL'.

3.5.4.3 Populations and pooled DNAs

The 'Pop_pool' model provides links to each of the DNA samples that constitute the population or pool, the markers that have been typed, the genotyping results, results of comparison of allele frequencies with those of other pools and information regarding inclusion of the population or pool in meta samples (Figure 3.17).

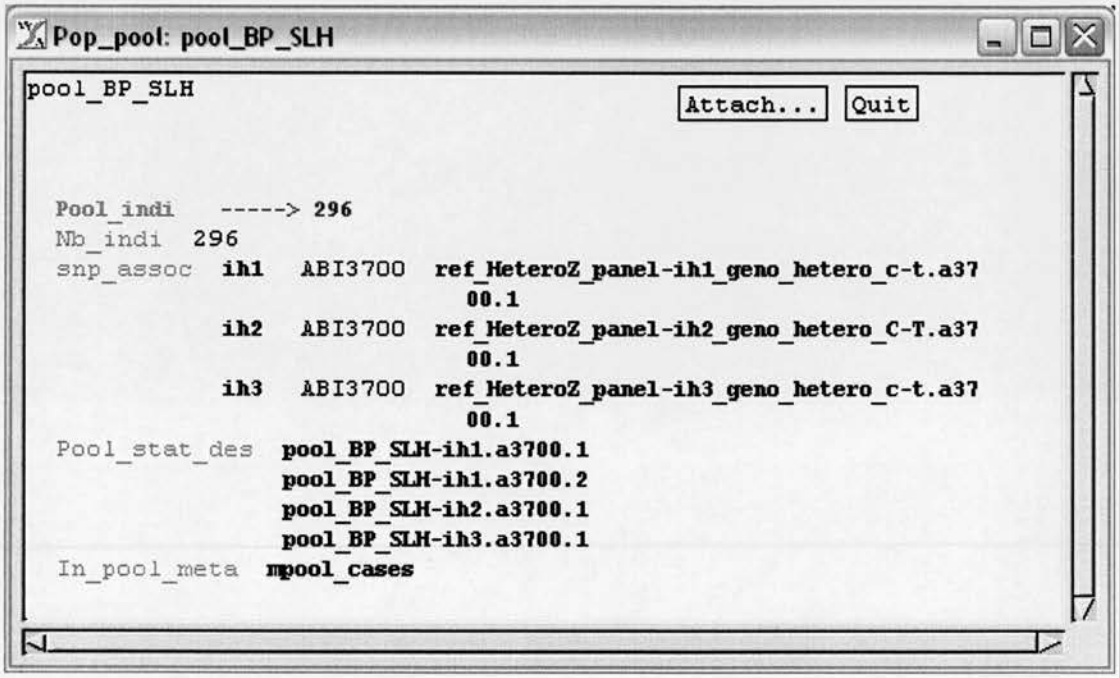


Figure 3.17: Management of pooled DNAs in ACeDB. Clicking on the tag 'Pool_indi' displays the list of links to individual samples in the DNA pool. The tag 'snp_assoc' lists the SNP loci for which the pool has been typed. For each of these markers, the genotyping platform is indicated and a link is provided to the data set for the heterozygote reference at that locus used to infer allele frequencies in this pool. The tag 'Pool_stat_des' provides links to data sets for genotyping assays at these SNP loci. The tag 'In_pool_meta' indicates that genotyping results for this pool have been merged with that for other pools forming the meta sample 'mpool_cases'.

3.5.4.4 Groups of populations or pooled DNAs

The 'Pop_pool_meta' model stores data on the sets of populations or pools combined; the markers typed on the items, the statistical description of the merged data sets, and information on the comparison of allele frequencies with those of other populations, pools, or meta samples (Figure 3.18).

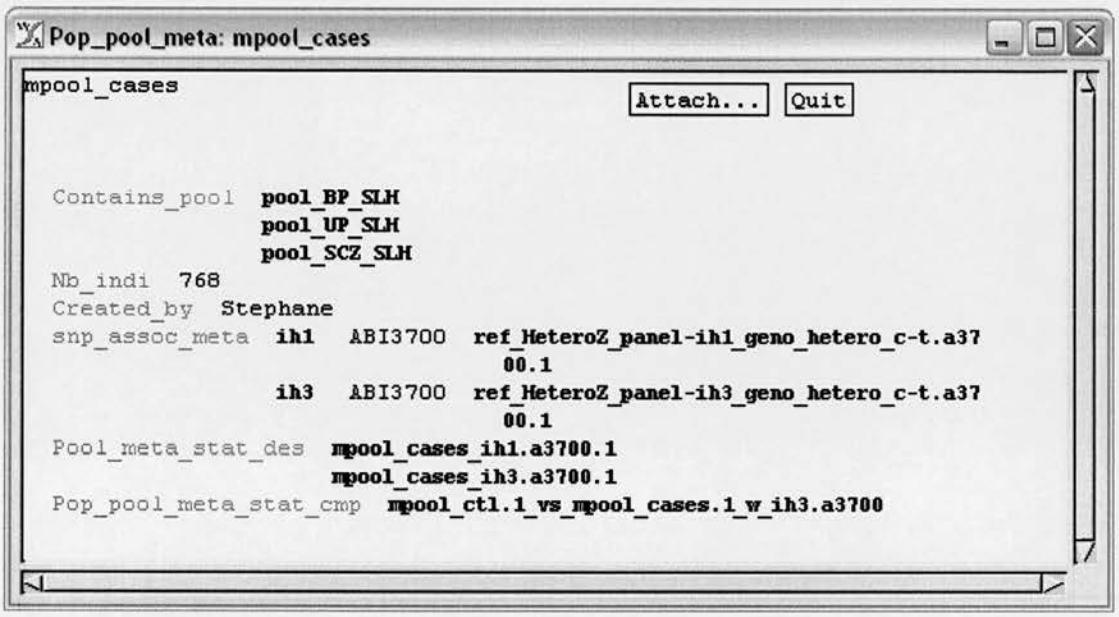


Figure 3.18: Management of meta sample of DNA pools in ACeDB. This example shows data stored for the meta sample of pools 'mpool_cases' which comprises three pools listed to the right of the tag 'Contains_pool'. The tag 'snp_assoc_meta' displays links to the marker studied and to the heterozygote references used to infer allele frequencies, accessible via the links of the tag 'Pool_meta_stat_des'. The tag 'pop_pool_meta_stat_cmp' provides links to 'Stat_cmp' objects containing data on the association studies involving this meta sample (compared here to the meta sample 'mpool_ctl').

3.5.4.5 Descriptive analysis of genotyping assays

The novel model 'List_Values' was created to store genotyping data of individual samples used in heterozygote references, as well as lists of peak height ratios and of allele frequencies in pools and groups of pools (Figure 3.19).

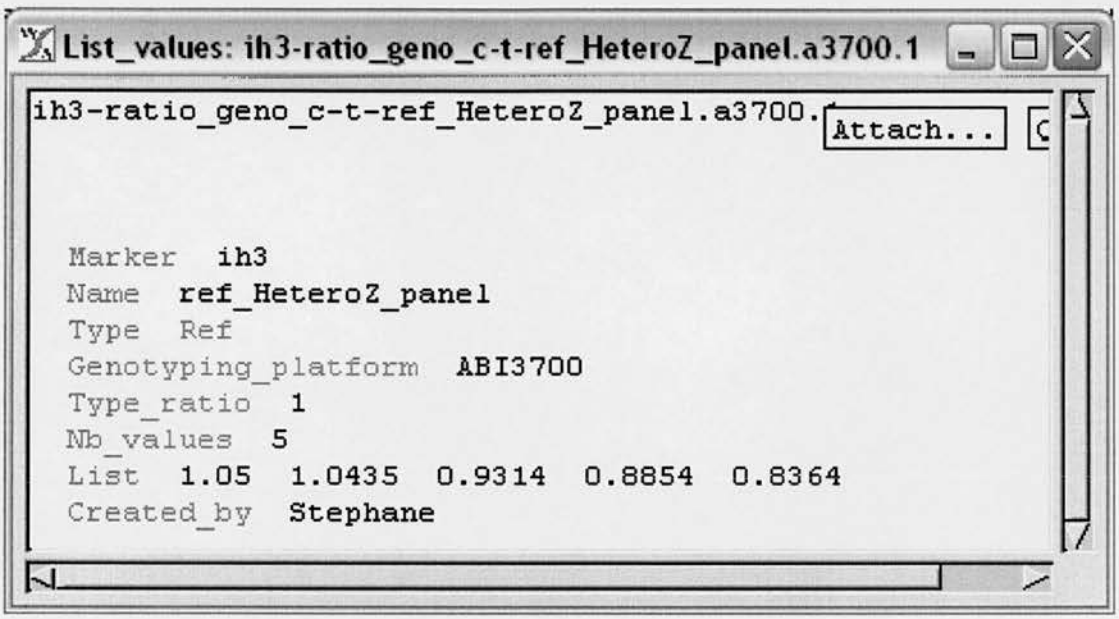


Figure 3.19: Management of genotyping data for heterozygote references in ACeDB. 'List_values' objects store sets of raw data. The example shown here shows the list of peak height ratios, displayed with the tag 'List', obtained for the SNP 'ih3' (tag 'Marker') with the heterozygote reference 'ref_HeteroZ_panel' (tags 'Name' and 'Type') using the 'ABI3700' platform (tag 'Genotyping_platform').

The statistical description model 'Stat_des' stores the results of the statistical analysis of genotyping data at a given marker for a given pop, pool or meta sample, such as allele frequencies (3.20). It also provides links to other relevant objects, such as the list of peak height ratios used to infer allele frequencies (Figure 3.20).

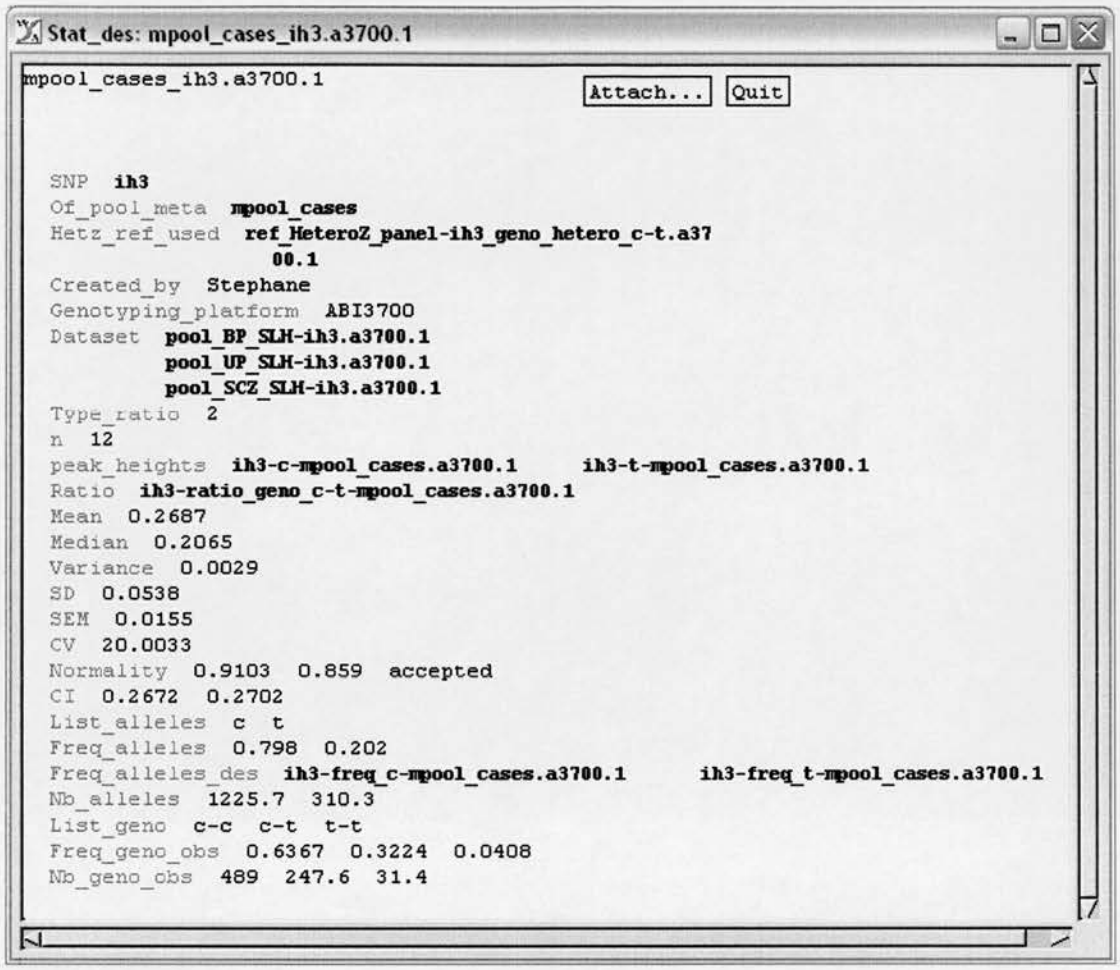


Figure 3.20: Management of genotyping results in ACeDB. 'Stat_des' object for results of the analysis of the meta sample 'mpool_cases' (tag 'Of_pool_meta') at the SNP 'ih3' (tag 'SNP'), using the heterozygote reference indicated by the tag 'Hetz_ref_used'. The object display links to the data sets that were merged to create the data set detailed here (tag 'Dataset'), to the object containing values of peak heights used (tag 'peak_heights') to derive peak height ratios (tag 'Ratio'). It also stores details on the peak height ratio (eg tag 'mean', Variance'), and numbers and frequencies of alleles. It also provides links to the 'Stat_des' objects for details on allele frequencies (tag 'Freq_alleles_des').

3.5.4.6 Allelic association studies

The model 'Stat_comp' (for 'statistical comparison') stores the results of classical and pooled DNA-based association studies (Figure 3.21).

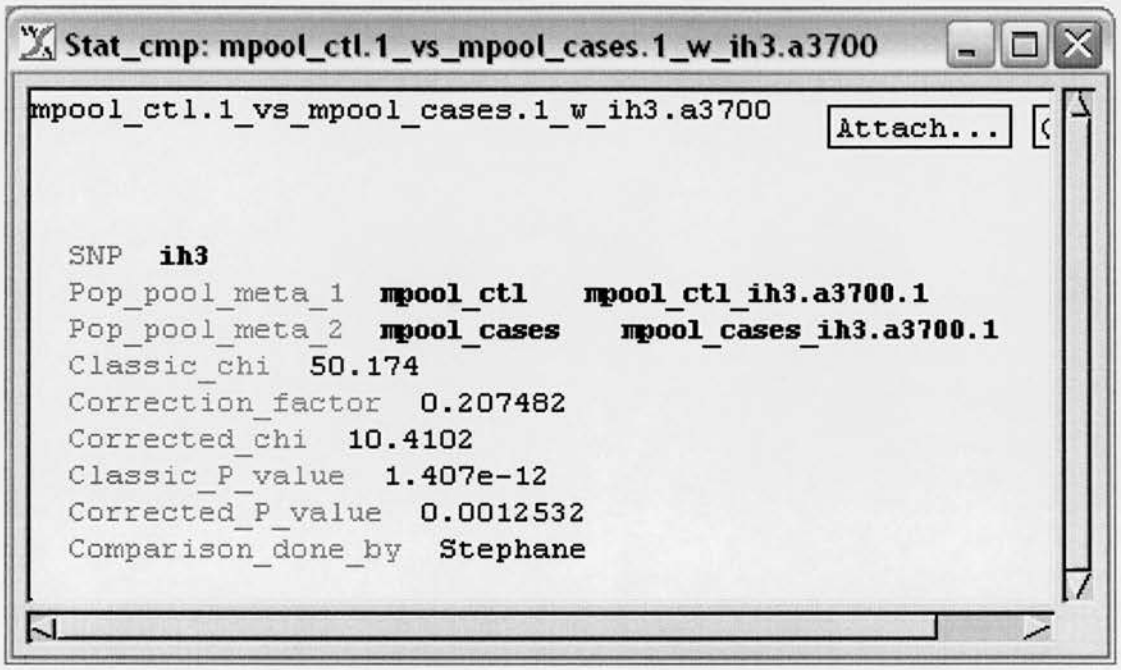


Figure 3.21: Management of allelic association data in ACeDB. Strings in bold are links to other relevant objects, such as the marker (ih3), the samples (here two meta samples of pools: 'mp_mood' and 'mpool_test'), or genotyping results used in the comparison of the samples (to the right of the samples names). The remaining data are test statistics, associated p-values, and in the case of pooled DNA, the correction factor used to account for error in the inference of allele numbers in pooled DNA.

3.6 Discussion

Large scale case-control association studies involve the genotyping of hundreds of markers in DNA pools or hundreds of individual samples and thus generate a large amount of data. No appropriate software existed to manage allelic association produced by the group and one was developed to address issues of data submission, analysis, storage, access, integrity and flexibility. The ACeDB database used in the laboratory was extended to allow storage of all information relevant to: i) genotyping assay experimental conditions and results, ii) DNA pool construction, iii) meta data sets, and iv) allelic association studies based on single DNA pools, single set of individual samples in a population, or merged data sets. This was achieved by modifying existing models and creating new ones. With these new models, one can now navigate from a marker to the pools or populations tested with that marker, to the data obtained, and to the results of the comparison of allele frequencies in that pool or population to those of other pools or populations. In addition, a user-friendly CGI front-end to the database was written to allow easy and accurate submission of new data by consulting the database and using a series of HTML forms, as well as their automatic examination, via descriptive statistical analysis and association studies.

For classical association studies, based on individual sample genotypes at a chosen marker, allele numbers in each selected population, or group of populations, are calculated from the genotypes of all individuals. A classical chi-square test is then performed on this data and its results are displayed. These are read into the database on demand. In the literature, most studies reporting the use of DNA pools in an association study perform a χ^2 test to detect the independence of the allele frequencies and phenotypes (Risch and Teng, 1998). On the other hand, allele numbers in pool of individual DNAs are not counted but must be inferred from peak height in heterozygotes and the pool itself. Once entered in a form, peak height ratios obtained for a set of heterozygote references or a set of single DNA pool replicates are run through a script that carries out a statistical analysis necessary for subsequent association studies (mean frequencies of the alleles, standard deviation, standard error). In association studies based on pools, the adjusted χ^2 test (T_{adj}) that accounts

for errors in inferring allele frequencies is used in place of the classical χ^2 test. The interface also allows the analysis of a data set produced by combining data from several populations or pools. A table of results is then displayed. The new data and their statistical description can then be read into the database.

DNA pooling aims to accelerate association studies by identifying SNPs of interest out of hundreds of markers. The conventional method of direct counting of alleles by genotyping individuals separately is then necessary to confirm the association of the highlighted SNP alleles with the disease. The interface for ACeDB therefore also allows the submission and analysis of DNA pool data and individual genotypes. As association studies are often hindered by the limited size of the samples used, the possibility of merging data sets from different phenotypes, or geographic location is also an advantage. The two tools described here are not specific to the diseases studied by the group and can potentially be used to support allelic association studies aiming to identify susceptibility loci of any disease. It may also be possible to extend the ACeDB database further to facilitate allelic association studies based on haplotypes.

The goal of an association study is to detect differences in allele frequencies between affected and unaffected people that are expected to indicate differences in the function or regulation of genomic sequence variants at the loci identified. Markers tested are either evenly distributed throughout a region or entire genome or chosen within candidate genes and putative or known regulatory elements. Once a set of candidate genes has thereby been refined, the next step in identifying susceptibility loci for a complex disease is to characterise these genes. An accurate structure of the genes and related sequences is in turn important to determine which markers should be tested in the association study. Annotation of functional elements, either known or predicted, in the 4p region is hence highly valuable.

CHAPTER 4

EVALUATION OF *IN SILICO* IDENTIFICATION OF *CIS*-REGULATORY SEQUENCES

4 EVALUATION OF *IN SILICO* IDENTIFICATION OF *CIS*-REGULATORY SEQUENCES

4.1 Introduction

A protocol was designed to predict Transcription Factor Binding Sites (TFBS) in upstream regions of genes by combining comparative genomics in human and mouse to prediction of known TFBS and prediction of novel motifs. This protocol was benchmarked against publicly available upstream regions of 28 human genes for which the function of TFBS had been verified experimentally (Dermitzakis and Clark, 2002). These sequences were masked for repeats and searched against a database of repeat-masked 2 Kb regions upstream of known genes in the mouse genome. Conserved human regions were selected on the length and percentage of identity of their alignment with mouse sequences. These conserved human sequences are referred to as 'CSSs' (for 'Conserved Sub-Sequence with single mouse sequence') and were searched for known TFBS in the TRANSFAC database. The subset of these human conserved regions which were matched by at least four mouse sequences, 'CSSm' (for 'Conserved Sub-Sequence with multiple sequences'), were also used together with the corresponding mouse regions to identify motifs over-represented in these sets of sequences. Accuracy of this protocol was assessed for each prediction programs separately, as well as for flexible combinations of these prediction programs.

4.2 Human-Mouse comparative genomics

4.2.1 Sequence conservation

Results of the comparative genomics analysis of upstream regions of the human genes containing experimentally verified *cis*-elements are shown in Table 4.1. Conserved regions larger than 50 bp and sharing more than 65% identity with mouse sequences were selected for further analysis. These sequences, 'CSSs', overlapped 64% of the length of the 28 searched sequences. These conserved human sequences were 153 bp long on average and their length ranged from 50 bp to 573 bp. The average percentage of identity shared between human and mouse sequence was 66.6% and ranged from 65% to 89%. These CSSs were used to predict known TFBS. Human conserved regions matched by at least four regions of mouse sequences were used together with the non-redundant matching sequences in mouse to identify motifs over-represented in these sets of sequences. Seventeen such sets, CSSm, were defined, for 11 genes (Table 4.1), and contained on average five sequences. The length of these CSSm was 52 bp on average and ranged between 50 bp and 58 bp. The total length of these conserved regions was 0.9 Kb, which represents 3% of the total length of the 28 human sequences.

4.2.2 Conservation of functional *cis*- regulatory elements

The location of each functional *cis*-element in the set of human sequences were compared to that of regions conserved between human and mouse to determine whether it had been conserved through evolution. Human sub-sequences showing at least 65% identity to the mouse sequence over at least 50 bp were chosen as conserved sequences (CSSs). Each *cis*-element was classified as conserved if at least 50% of its length was overlapped by a conserved sequence. Out of a total of 100 functional human TFBS, 79 were conserved in mouse.

Sequence		CSSs		CSSm		Regions				motifs			
Name	lgth	nr	lgth	nr	lgth	nr	Lgth	scon	mcon	nr	lgth	scon	mcon
ABCC2	3000	9	1007	0	0	1	234	1	0	1	14	1	0
AGT	590	2	561	0	0	2	42	2	0	5	41	4	0
AKR1C1	1132	4	957	1	51	1	581	1	0	7	58	7	0
ALB	1286	7	1145	2	107	1	37	1	1	1	37	1	1
ALDH2	721	4	498	0	0	1	602	1	0	3	74	2	0
ALDOB	200	1	162	0	0	1	200	1	0	2	39	1	0
APOB	1980	11	1481	1	51	3	277	3	0	8	157	5	0
APOCIII	1435	5	1056	2	110	1	857	1	0	2	22	1	0
C4BPB	577	1	573	0	0	1	201	1	0	2	14	2	0
CFTR	843	3	664	0	0	1	843	1	0	11	615	7	0
CYP1A2	3293	4	955	1	53	1	259	0	0	1	7	0	0
CYP7A1	200	1	198	1	50	1	200	1	0	4	49	4	1
F10	814	3	505	1	52	1	197	1	0	2	65	2	0
F7	1017	2	570	0	0	1	201	1	0	2	45	2	0
G6PC	200	1	197	1	53	1	200	1	0	10	74	10	3
HBB	204	1	191	0	0	1	204	1	0	7	91	6	0
HBE1	168	1	167	0	0	1	168	1	0	6	129	6	0
HMGCR	1227	9	899	0	0	1	200	1	0	1	16	1	0
IGF1	200	1	191	0	0	1	200	1	0	2	26	2	0
IGF2	1523	7	1286	1	53	1	22	1	0	1	9	1	0
INS	200	1	194	2	101	1	200	1	1	1	15	1	0
LIPC	1728	4	689	0	0	2	433	1	0	3	21	0	0
PAH	200	1	197	0	0	1	200	1	0	2	53	2	0
PRL	3300	15	2070	4	211	1	905	1	0	1	21	1	0
PROC	200	1	173	0	0	1	200	1	0	8	60	8	0
SEPP1	1812	10	1289	0	0	1	201	0	0	2	21	0	0
SLC2A2	1823	11	1037	0	0	1	46	0	0	3	24	0	0
UGT1A1	200	1	198	0	0	1	200	1	0	2	23	2	0
sum	30073	121	19110	17	892	32	8110	28	2	100	1820	79	5

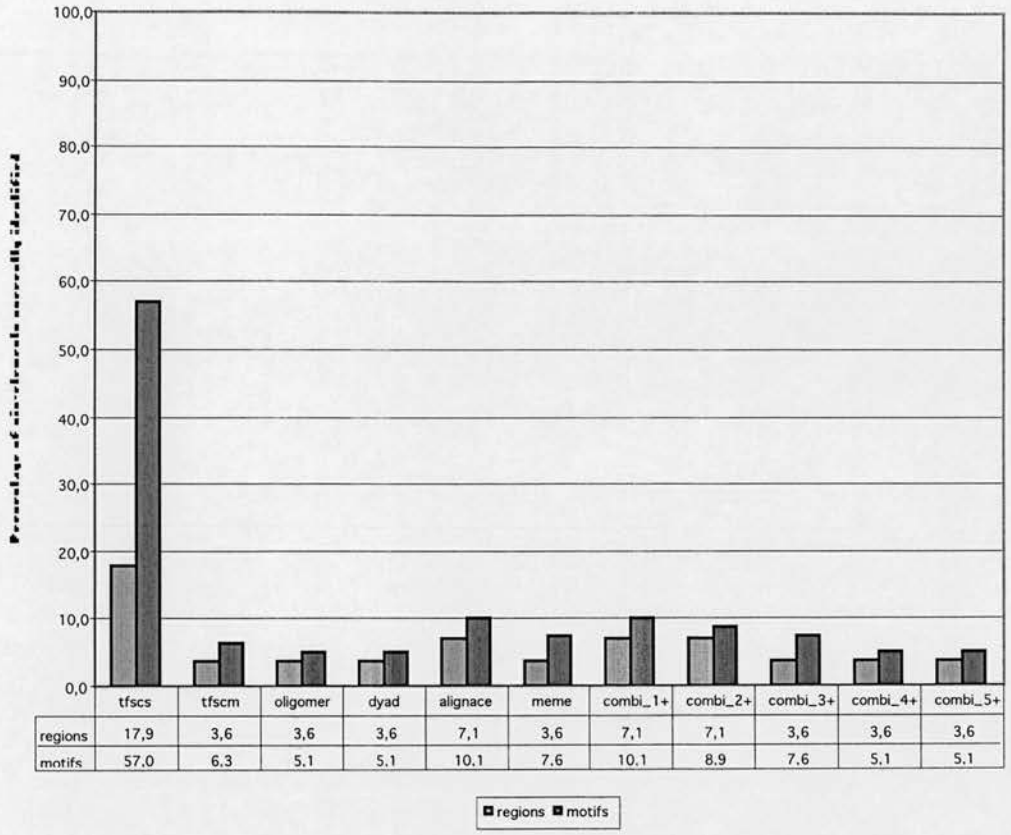
Table 4.1: Conservation in mouse gene upstream flanking regions of human sequences with known functional *cis*- regulatory elements. Data were obtained by searching the repeat-masked non-coding sequence of each of the 28 human genes listed that contain functional *cis*- regulatory elements to repeat-masked 2 Kb sequences upstream of the translation start site of genes in the mouse genome retrieved from the Ensembl web site. Out of all conserved regions, regions longer than 50 bp and sharing more than 65% identity with the mouse sequences were included in the analysis. Lgth: total length in pb; nr: number of elements: CSSs : Conserved Sub-Sequences with Single mouse sequence; CSSm: Conserved Sub-Sequence with Multiple mouse sequences; regions: region required for promoter activity; motifs: regulatory motifs, which are located within the 'regions'; scon: number of *cis*-elements located in CSSs regions; mcon: number of *cis*-elements located in CSSm regions.

4.3 Detection of functional cis- regulatory elements

4.3.1 Accuracy of individual prediction methods

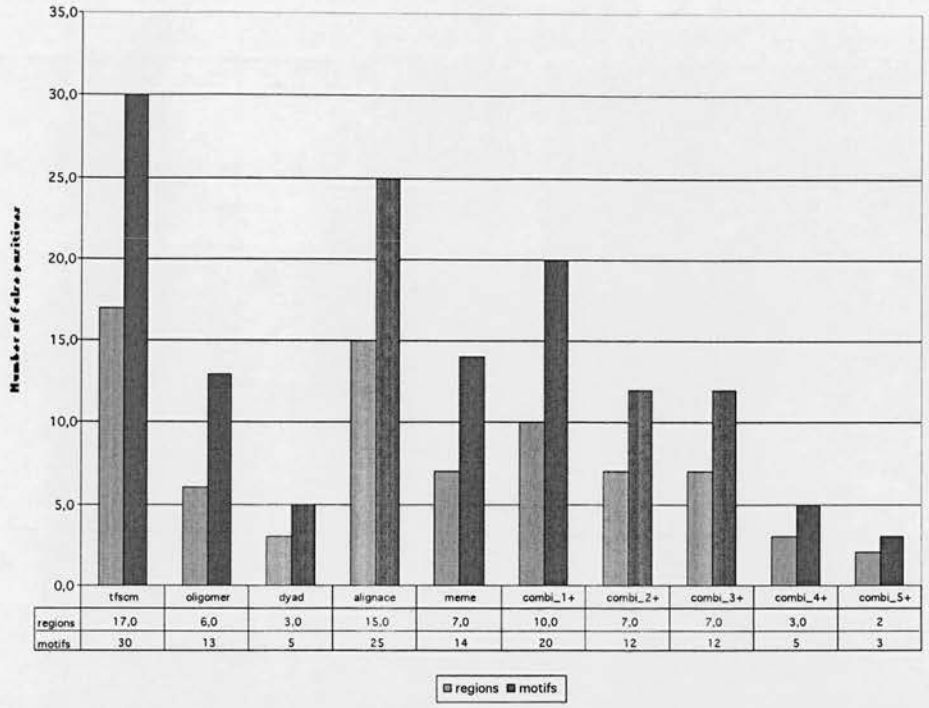
Prediction of known functional *cis*-elements was performed on human regions conserved in at least one mouse sequence, 'CSSs', using the EMBOSS tool tfscan to search the TRANSFAC database. Total number and total length of these conserved regions are displayed in the third and fourth columns of Table 4.1. Prediction of novel motifs was carried out by identifying motifs over-represented in sets of sequences each containing the human conserved region and the mouse sequences matching this region ('CSSm'). Total number and total length of these conserved regions are displayed in the fifth and sixth columns of Table 4.1.

Searching TRANSFAC for known motifs in CSSs regions ('tfscs' in figure 4.1) was the most sensitive approach as 18% of regions and 57% of motifs were matched by motifs predicted by tfscan. However, searching TRANSFAC for known motifs in CSSm regions ('tfscm' in figure 4.1) performed poorly, with only 4% and 6% of regions and motifs identified respectively. Low sensitivity was also observed with the other programs tested. Only 7% of regions and 10% of motif were identified by motifs predicted by AlignACE, which performed best out of the four programs for novel motif discovery. Prediction methods differed in the number of false positives (predicted motifs that do not match functional regulatory regions or motifs) generated. Prediction by tfscan on CSSs generated 387 and 528 false positives for 'regions' and 'motifs' respectively. With CSSm, the highest number of false positives, for regulatory regions and motifs, was produced by tfscan, followed by AlignACE, MEME, Oligo-analysis and Dyad-analysis (Figure 4.2). However, all programs performed equally badly in terms of false positive rates (Figure 4.3). Indeed all programs over-predicted massively, with a rate of false positives for each method higher than 70%.



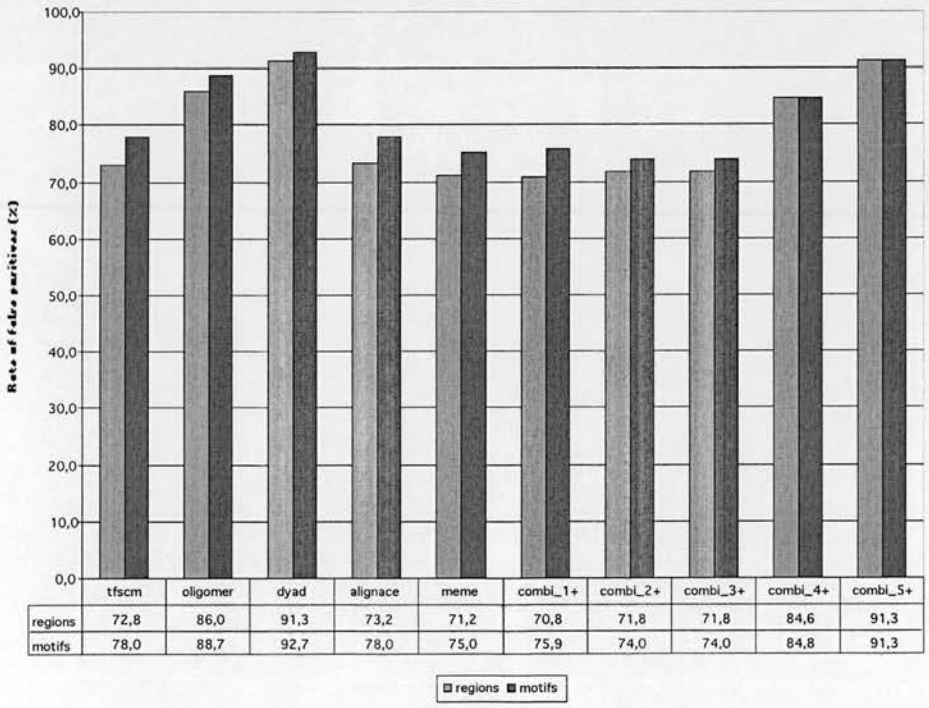
	tfscs	tfscm	oligomer	dyad	alignace	meme	combi_1+	combi_2+	combi_3+	combi_4+	combi_5+
regions	17,9	3,6	3,6	3,6	7,1	3,6	7,1	7,1	3,6	3,6	3,6
motifs	57,0	6,3	5,1	5,1	10,1	7,6	10,1	8,9	7,6	5,1	5,1

Figure 4.1: Percentage of known functional *cis*-elements detected with the method combining human-mouse comparative genomics to motif prediction. Each functional *cis*-element was classified as detected if it was overlapped by predicted motifs by more than 50%. The percentage of correctly identified *cis*-element was computed for both type of *cis*-elements, regulatory regions, required for the activity of the promoter ('regions'), and known sequence motifs within these regions ('motifs') for each prediction program, and for flexible combinations of these programs. 'tfscs': tfscan ran on CSSs; 'tfscm': tfscan ran on 'CSSm'; 'oligomer': Oligo-analysis; 'dyad'; Dyad-analysis; 'alignace': AlignACE; 'meme': MEME; 'combi_1+': any one of the following five methods: 'tfscm', 'oligomer', 'dyad', 'alignace' and 'meme'; 'combi_2+': any two of the above mentioned five methods; 'combi_3+': any three of the above mentioned five methods; 'combi_4+': any four of the above mentioned five methods; 'combi_5+': all five methods.



	tfscm	oligomer	dyad	alignace	meme	combi_1+	combi_2+	combi_3+	combi_4+	combi_5+
regions	17,0	6,0	3,0	15,0	7,0	10,0	7,0	7,0	3,0	2
motifs	30	13	5	25	14	20	12	12	5	3

Figure 4.2: Number of false positives generated by the method combining human-mouse comparative genomics to motif prediction. Details are as in figure 4.1.



	tfscm	oligomer	dyad	alignace	meme	combi_1+	combi_2+	combi_3+	combi_4+	combi_5+
regions	72,8	86,0	91,3	73,2	71,2	70,8	71,8	71,8	84,6	91,3
motifs	78,0	88,7	92,7	78,0	75,0	75,9	74,0	74,0	84,8	91,3

Figure 4.3: Rate of false positives generated by the method combining human-mouse comparative genomics to motif prediction. Details are as in figure 4.1.

4.3.2 Flexible combination of motif prediction methods

In an attempt to reduce the number and rate of false positives, clusters of overlapping motifs were defined and these were used in place of individual motifs. Categories of motif clusters used to assess identification of *cis*-elements were derived based on a minimal number of programs that predicted motifs contained by these clusters, i.e. clusters made of motifs predicted by at least one program, clusters made of motifs predicted by at least two programs and so on. This approach also aimed to test whether using regions overlapped by motifs predicted by at least a certain number of methods would be as sensitive, and also more specific than the approach based on individual prediction methods.

Figure 4.1 shows the proportion of functional *cis*-element identified by predictions generated by combinations of at least a certain number of prediction programs, e.g. 'combi_2+' for the proportion of functional *cis*-element identified by clusters of overlapping predicted motifs from at least two motif prediction programs. The number of motif clusters containing motifs predicted by at least one method was similar to that obtained with AlignACE alone. Sensitivity obtained with clusters of motifs predicted by at least one method was 7% for functional regions and 10% for functional motifs (Figure 4.1). As expected, the percentage of functional regions and motifs identified decreased as the minimum number of methods in the combination increased. With clusters of predicted motifs from all five programs, sensitivity dropped to 4% for functional regions and to 5% for functional motifs, respectively. The number of false positives also decreased as the minimum number of programs represented in these clusters increased (Figure 4.2). However, the rate of false positives for any flexible combination was still higher than 70% (Figure 4.3).

4.4 Discussion

A computational protocol was designed to predict putative regulatory sites in non-coding regions upstream of genes. It combined the comparative genomics approach applied to non-coding sequences upstream of human and mouse genes, and the prediction of known TFBS and novel over-represented motifs. The accuracy of this protocol was benchmarked against a set of regions upstream of human gene known to contain functional *cis*-elements and which were available publicly (Dermitzakis and Clark, 2002). These human sequences were masked for repeats and aligned to repeat-masked 2 Kb non-coding sequences upstream of the transcription start site of Ensembl known genes in mouse. Conserved human sequences were selected on their length and degree of similarity to the mouse sequences. These non-coding conserved upstream sequences were used to predict known TFBS and to identify over-represented motifs. The accuracy of the prediction was assessed by comparing the location of the known functional *cis*-elements with that of the motifs predicted. Selecting conserved regions showing similarity higher than 65% over at least 50 bp, 79 out of 100 functional sites were conserved between human and mouse. This is consistent with the observation that a third of the human functional sites are not functional in rodents (Dermitzakis and Clark, 2002).

The accuracy of the prediction of known and over-represented motifs varied between programs. The highest number of *cis*-elements identified was obtained by searching TRANSFAC in sub-sequences conserved in at least one mouse upstream region. However, 42% of the functional sites were not identified by predicted known motifs. This might be due to the absence of some of these functional sites in TRANSFAC, and the use of tfscan, which is based on consensus sequences of TFBS, and was run without allowing mismatches. Prediction of known TFBS may have been more sensitive based on position weight matrices (PWMs). It may therefore prove valuable to use programs based on PWMs to identify matches to known TFBS, such as MATCH (Kel et al., 2003) or MatInspector (Quandt et al., 1995). However, the use of PWMs requires the choice of a score threshold to select meaningful matches. Unfortunately such threshold can rarely be defined based on experimental data because there is not enough information so far to do so systematically. Score thresholds are therefore usually arbitrary (Werner, 2000). Although sensitivity may

be increased by using PWMs, the rate of false positives would likely remain relatively high. The lower sensitivity of programs that identify over-represented motifs may be due to the fact that the number of occurrences of functional sites in the set of sequences may be insufficient for these motifs to be identified as over-represented. The number of motifs predicted varied between motif prediction programs. This variation in the number of motifs predicted may reflect the type of motifs modelled by different programs, and the options used to run these programs. For example, the number of motifs predicted by dyad-analysis was high because it predicted tri-nucleotide pairs separated by a spacer of length varying between 1 and 20. Oligo-analysis was run to predict oligomers of 6 bp only. Running oligo-analysis to predict motifs of varying size would not necessarily improve sensitivity but would likely result in more false positives. Also, as the highest scoring motifs predicted by any methods are not necessarily motifs with a biological meaning (Werner, 2000), selecting the top-scoring motifs predicted in each set of conserved sequences may reduce the number of motifs considered but also decrease sensitivity. In addition, the high rates of false positives observed are higher than those reported by studies using these programs (Werner, 2000), indicating that the approach applied to this set of sequences may not be optimal.

In order to limit the number of predicted motifs, overlapping predicted motifs were grouped into clusters containing motifs predicted by different methods. This approach proved efficient in eliminating the redundancy of overlapping motifs. In addition, classifying these clusters according to the number of methods that generated motifs making up the clusters addressed two questions. First, would selection of predictions corresponding to a flexible combination of motif prediction programs maintain sensitivity despite setting more stringent criteria, or increase sensitivity by joining motifs from different programs into a motif larger than the initial single ones? Second, would it also increase specificity, by giving more weight to regions matched by several predictions. The number of functional *cis*-elements identified by cluster containing predictions of at least one method was the same as that identified by motifs predicted by AlignACE. Sensitivity obtained by selecting clusters of predictions generated by more programs decreased as the required number of programs increased. Although the number of false positives reduced by

considering flexible combination of methods, the rate of false positives was as high as that obtained by considering motif prediction programs separately.

The protocol tested aimed to annotate putative functional *cis*-elements in non-coding regions upstream of genes in a region linked to psychosis. It was designed to help characterise these genes and provide clues on nucleotide variation in putative regulatory sites in promoters. Bench scientists would then test these polymorphisms for association to the illness. This protocol was tested on a set of promoter regions of genes containing known functional *cis*-elements and proved to be partially successful. Indeed, up to 60% of the functional sites conserved between human and mouse could be identified. However, all motifs prediction programs generated very high rates of false positives. Flexible combinations of programs were used in place of individual, redundant predicted motifs to try and maintain the same sensitivity and increase specificity. Although this approach reduced the number of false positives, the rate of false positives was not reduced to a satisfactory level. In addition, sensitivity was reduced by half in the most flexible case (clusters of predictions from at least two programs). The protocol tested here was therefore not suitable for predicting *cis*-elements to help characterise genes in the 4p region linked to psychosis. These results also showed that another approach was necessary. Multi-species comparative genomics has been reported to help address the problem of annotating putative functional sequences (Thomas et al., 2003). Non-coding regions conserved between several species representing a range of evolutionary distances are indeed more likely to be functional than regions conserved between human and mouse, as the latter contain functional and neutral sequences. The approach used to annotate putative functional sites in non-coding regions conserved across several species is the subject of the next chapter.

CHAPTER 5

EVALUATION OF SCORING SCHEMES FOR COMPARATIVE GENOMICS WITH VERTEBRATE SPECIES

5 EVALUATION OF SCORING SCHEMES FOR COMPARATIVE GENOMICS WITH VERTEBRATE SPECIES

5.1 Introduction

Comparative genomics of the human and mouse sequences has been reported to be an efficient tool to identify coding sequences and functional non-coding regions such as gene regulatory elements (Tagle et al., 1988; Pennacchio and Rubin, 2001). Chapter 4 reported results of the application of this approach, which was combined to the prediction of known and novel motifs, to a test set of known functional *cis*-regulatory elements. The protocol designed generated a number of false positives too high to provide a limited number of sequences to bench scientists to guide them in the identification of putative non-coding disease-associated polymorphisms. Comparing genomic sequences from more than two species has the advantage over studying two species of reducing the amount of conserved neutral sequences, as divergence eliminates these sequences differently in each species and lineage (Duret and Bucher, 1997; Thomas et al., 2003). The present chapter aims to assess the ability of comparative genomics applied to the currently available whole-genome sequences from vertebrate species of detecting non-coding functional sequences. Four multiple alignment scoring schemes, which are described in chapter 2, were benchmarked against a publicly available set of experimentally verified functional regulatory regions (Elnitski et al., 2003), identifying the most useful scoring scheme.

The first system was the simple mean of the non-human species' pairwise percentage of identity to the human sequence. The second and third systems involved weighting the pairwise percentage of identity to the human sequence. The second system was based on a simple weight reflecting the relative evolutionary distances between human and the other species, that is 1 for the rodent sequences, 3 for the chicken sequences and 4.5 for the fish sequences, since the evolutionary distance between human and rodent is approximately 100 millions years, that between human and chicken is approximately 300 millions years and that between human and fish is approximately 450 millions years. For the third measure, the weight was the estimate of synonymous substitutions, dS, for each species, or 1 if dS was greater than 1. The fourth scoring system is derived from the probability of observing the measured amount of conservation given the estimate of the rate of neutral evolution, dS, of the gene in the species being compared to human.

5.2 Sequence conservation in rodents, chicken and fish

Four sets of genes with one or more experimentally verified regulatory region were considered based on the species in which an ortholog of the human gene was identified: i) 'HMR' (human, mouse and rat), 'HMRC' (human, mouse, rat and chicken), 'HMRZF' (human, mouse, rat, zebrafish and Fugu), 'HMRCZF' (human, mouse, rat, chicken, zebrafish and Fugu). Pairwise rate of synonymous substitution dS was estimated for each species compared with human. In mouse and rat, 20% of the genes are saturated with synonymous substitutions (Appendix 9.1.1). All genes in the chicken and in the fish show saturation, as expected. Figure 5.1 shows the average level of conservation of known functional elements and sequences with no known function for each set. As expected, coding sequences appear to be more conserved on average than any other feature, while UTRs show an intermediate level of conservation, lower than that of the coding region, but higher than that of the untranscribed sequences. These trends are observed in all four sets (Figure 5.1). Levels of conservation obtained with the HMR set for the various annotation features are all higher than those obtained with the other sets as expected. Also, it should be noted that the average level of conservation of each type of annotation feature is highly variable. Figure 5.2 shows the distribution of the averaged percentage of identity and of the score based on the binomial probability as measured for a window sliding along the multiple alignments for the HMRC set. For all sets, the distribution of the window scores obtained with the scoring scheme based on the binomial probability, for the coding sequences, the regulatory regions and the non-coding sequences with no known function spreads across the range of possible values. It is skewed towards low values for the non-coding sequences and regulatory regions and towards high values for the coding sequences. This difference in the shape of the distributions is accentuated when chicken and fish sequences are considered (data not shown). The distribution for the regulatory regions has a similar shape to that of the non-coding region and is slightly shifted to higher values, reflecting a better conservation of the regulatory regions compared to the whole of the non-coding sequences. Also, its distribution overlaps that observed for the coding sequences, showing that some regulatory regions are at least as well conserved as some coding sequences.

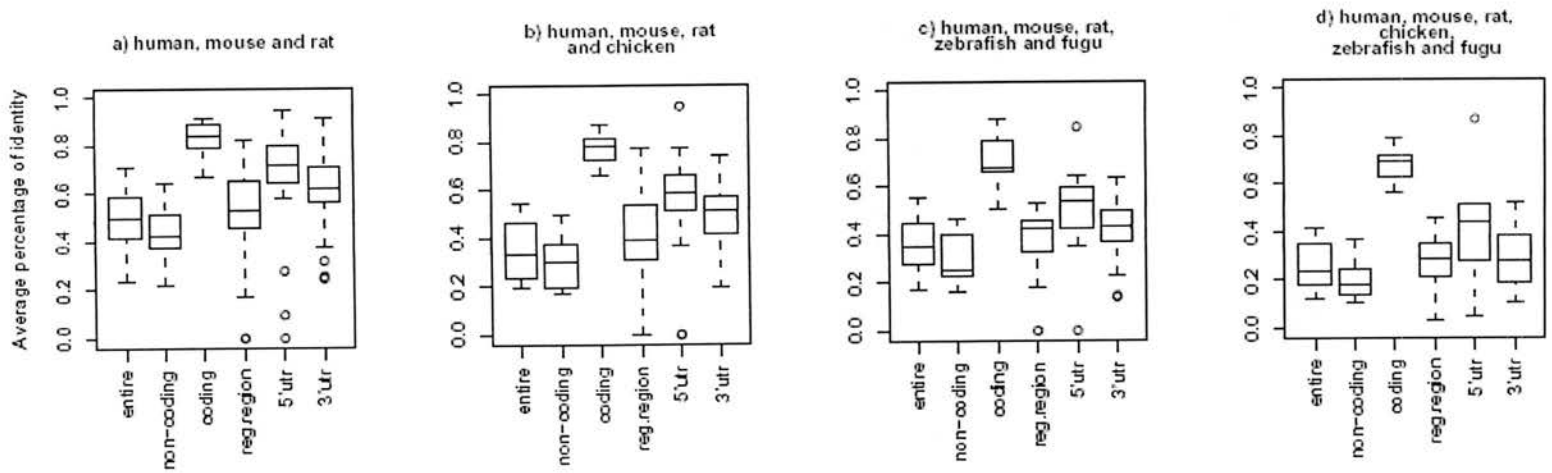
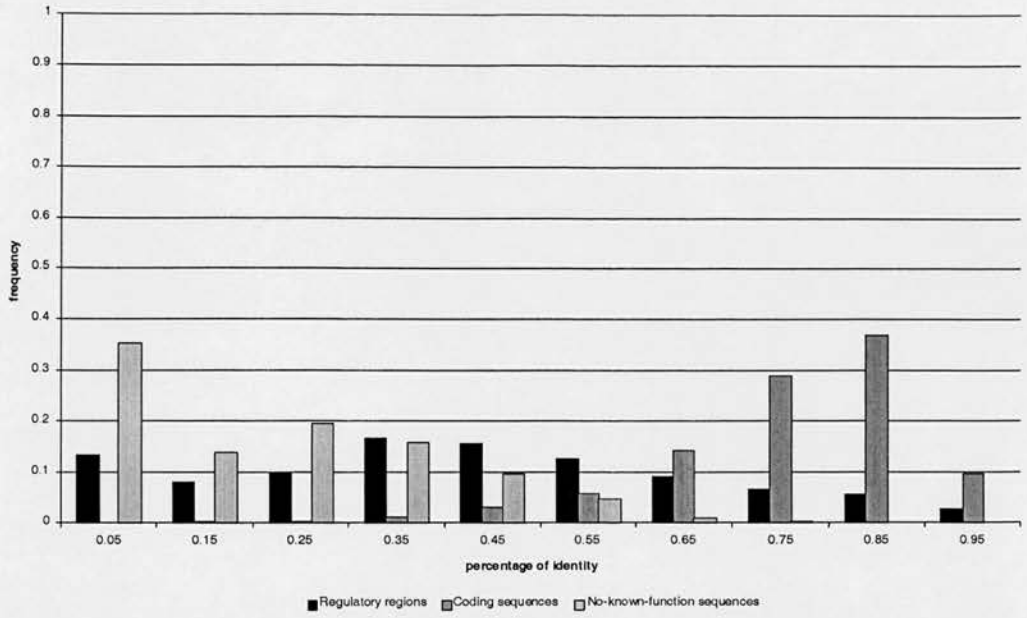


Figure 5.1 : Conservation of human genomic sequences in Vertebrates. Boxplots show the percentage of identity for several annotation features on the human sequence average over genes from four sets of orthologs: HMR (a), HMRC (b), HMRZF (c) and HMRCZF (d). Names of sets are explained in the text. 'entire': entire sequence; 'non-coding': non-coding sequences; 'coding': coding sequences; 'reg.region': regulatory region, 5'utr: 5' untranslated region; 3'utr: 3' untranslated region.

a) percentage of identity



b) 'binomial' score

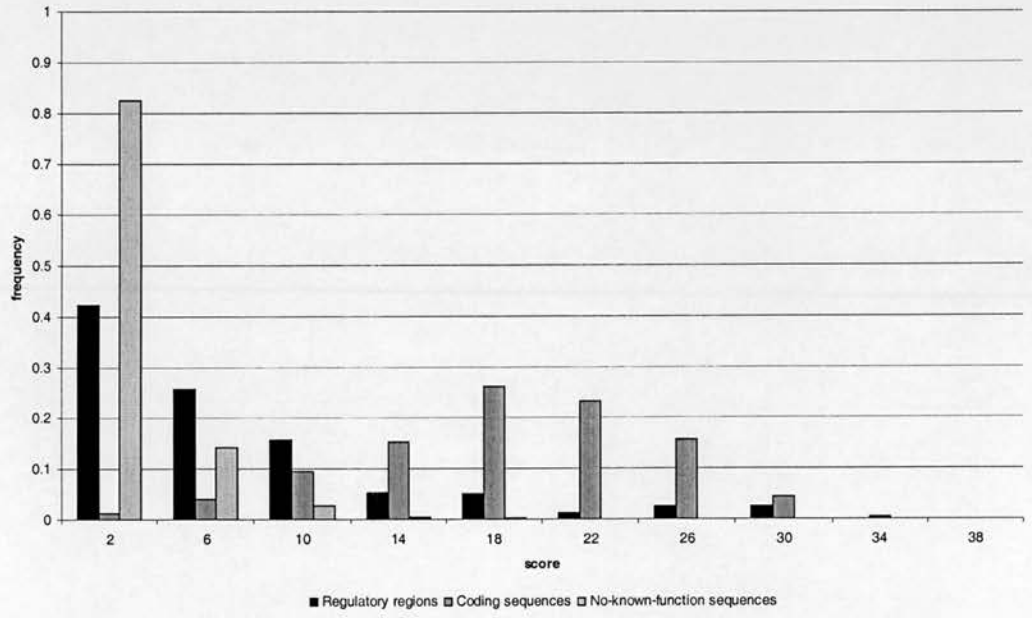


Figure 5.2: Distribution of window scores. The charts show the distribution of the average percentage of identity (a) and of the score based on the binomial probability (b) obtained for a 40 bp window sliding along the alignment of sequences from human, mouse, rat and chicken by 20 bp at time, for the known coding and regulatory regions, and the sequences with no known function.

5.3 Accuracy of detection of functional sequences

Table 5.1 shows the number and length of conserved sequences identified by each scoring scheme in each set of genes. Accuracy of detection of known functional sequences was similar for all four scoring systems, with all four sets (Table 5.2 and Table S2-S4 in Appendix 9.1.2). Accuracy obtained with the binomial score was slightly higher than that observed with simple scoring schemes when the sequences from fish were added. In spite of the similarity of accuracy between the scoring systems, the score based on the binomial probability was as good as or better than the other scoring systems across the data sets. Based on these data, the system of choice would therefore be the score based on the binomial probability. With the HMR set, using a window length of 40 bp and selecting regions that make up the percentage of the human sequence represented by coding sequences, 31 genes out of 40 (78%) contained regulatory regions that were detected. At the nucleotide level, the average sensitivity was 78% for the coding sequences and 34% for the regulatory regions (range 7-78%), considering only the genes for which the regulatory region was detected. The average specificity was 97% (range 89-100%). The sensitivity, specificity and positive predictive value of detection of sequence with a known function, that is coding and regulatory regions, were 60%, 97% and 75%, respectively. Accuracy was similar with the other sets (Table 5.2), although it is worth noting that fewer genes were analysed as orthologs could not be readily identified in chicken and fish for all human genes in the HMR set. It is noteworthy that regulatory regions could be identified in the majority of the genes. The inclusion of the chicken and fish sequences resulted in a slight increase in the sensitivity of detection of coding sequences (82% on average) and a slight decrease of the sensitivity of detection of regulatory regions (26% on average). Using a smaller window length of 20 bp resulted in identifying regulatory regions in up to four more genes, although the overall sensitivity and positive predictive probability were lower than with a 40 bp window, for coding and regulatory regions (Table 5.3). Using a 100 bp window resulted in the identification of fewer regulatory regions (Table 5.3). Figure 5.3 shows the conservation of the human *G6PC* gene sequence, the known annotation and the selected conserved regions.

		20 bp		40 bp		100 bp	
Set	score	count	length	count	length	count	length
HMR	bino	2485	43	1119	91	591	173
	diverg	2723	43	1237	92	655	174
	disty	2800	43	1252	92	657	175
	pid	2800	43	1252	92	657	175
HMRC	bino	906	65	543	109	320	186
	diverg	1173	60	682	103	394	181
	disty	1208	58	690	103	401	179
	pid	1216	59	688	103	396	181
HMRZF	bino	1156	56	663	98	396	166
	diverg	1165	56	704	93	396	168
	disty	1186	55	721	92	402	165
	pid	1194	56	712	94	396	168
HMRCZF	bino	683	62	434	98	248	174
	diverg	744	58	454	96	253	173
	disty	809	54	486	90	255	172
	pid	742	59	443	99	249	177

Table 5.1: Selected conserved sequences. The number and average length of selected conserved sequences obtained for three window lengths (20 bp, sliding by 10 bp, and 40 and 100 bp, sliding by 20 bp) are given for each set and each scoring system. Names of sets are as in the text. 'score': scoring system; 'count': number of selected conserved sequences; 'length': average length of selected conserved sequences across all genes in each set; 'bino': binomial score; 'diverg' score based on dS estimate, 'disty': score based on the distance in years between species; 'pid': score defined as the average percentage of identity.

length	set	score	sensitivity			specificity			ppv		
			rr	cds	fct	rr	cds	fct	rr	cds	fct
20 bp	HMR	bino	30	70	57	96	96	96	26	66	69
		diverg	30	70	57	96	96	96	26	66	69
		disty	30	71	58	96	96	96	26	66	69
		pid	30	71	58	96	96	96	26	66	69
	HMRC	bino	21	81	60	98	98	98	33	76	79
		diverg	20	80	60	98	98	98	31	76	78
		disty	21	80	59	97	98	98	30	74	77
		pid	21	81	60	98	98	98	32	75	77
	HMRZF	bino	22	79	65	97	98	98	28	78	80
		diverg	22	78	63	97	97	97	29	75	77
		disty	21	76	62	96	97	97	27	73	75
		pid	22	80	65	97	97	97	27	76	78
	HMRCZF	bino	26	81	66	98	98	98	40	78	81
		diverg	22	79	63	98	98	98	35	75	78
		disty	19	77	61	97	98	98	29	73	76
		pid	25	80	65	98	98	98	37	76	79
40 bp	HMR	bino	1	6	4	1	1	1	11	8	7
		diverg	1	6	4	1	1	1	10	8	7
		disty	1	6	4	1	1	1	11	8	7
		pid	1	6	4	1	1	1	11	8	7
	HMRC	bino	-1	1	0	0	0	0	1	0	0
		diverg	0	2	1	0	0	0	4	1	0
		disty	2	2	1	1	0	0	6	1	1
		pid	0	1	0	0	0	0	3	1	1
	HMRZF	bino	3	3	1	0	0	0	3	1	0
		diverg	4	2	1	0	0	0	7	1	0
		disty	0	2	1	0	0	0	6	1	1
		pid	3	2	1	0	0	0	4	1	0
	HMRCZF	bino	-2	-2	-3	0	-1	-1	-13	-5	-5
		diverg	-1	-1	-2	0	-1	-1	-12	-4	-4
		disty	-1	-1	-2	0	0	0	-10	-3	-4
		pid	0	-2	-3	0	-1	-1	-12	-4	-4
100 bp	HMR	bino	6	10	4	1	1	1	11	8	7
		diverg	7	10	4	1	1	1	11	9	7
		disty	6	9	4	1	1	1	12	9	7
		pid	6	9	4	1	1	1	12	9	7
	HMRC	bino	10	-5	-4	0	-1	-1	3	-8	-8
		diverg	9	-5	-4	0	-1	-1	2	-7	-8
		disty	7	-4	-4	1	-1	-1	2	-7	-7
		pid	10	-5	-4	0	-1	-1	3	-6	-7
	HMRZF	bino	8	-2	-3	0	-1	-1	3	-5	-6
		diverg	6	0	-1	0	0	0	1	-2	-2
		disty	2	-1	-2	0	0	0	5	-3	-3
		pid	6	-2	-3	0	0	0	4	-2	-3
	HMRCZF	bino	4	-9	-10	-1	-2	-2	-13	-13	-13
		diverg	15	-8	-8	0	-2	-2	-2	-11	-12
		disty	19	-7	-7	1	-1	-1	3	-9	-10
		pid	7	-8	-10	0	-2	-2	-9	-12	-13

Table 5.2: Accuracy of detection of functional sequences. Accuracy of detection was assessed at the nucleotide level for three window lengths: 20 bp, sliding by 10 bp, and 40 and 100 bp, sliding by 20 bp. Values shown are averages over all genes in each set. Names of sets are as in the text. Values displayed for windows of 40 and 100 bp are the difference between the values obtained with these windows and that obtained with the 20 bp window. 'bino': binomial score; 'diverg' score based on dS estimate, 'disty': score based on the distance in years between species; 'pid': score defined as the average percentage of identity; 'rr': regulatory region; 'cds': coding sequence; 'fct': functional sequence (i.e. 'rr' or 'cds'); ppv: positive predictive value.

			20 bp		40 bp		100 bp	
Set	score	nrg	nrpg	%pg	nrpg	%pg	nrpg	%pg
HMR	bino	40	34	85	30	75	16	40
	diverg	40	34	85	30	75	16	40
	disty	40	34	85	30	75	16	40
	pid	40	34	85	30	75	16	40
HMRC	bino	22	16	73	14	64	9	41
	diverg	22	17	77	14	64	9	41
	disty	22	16	73	12	55	7	32
	pid	22	17	77	14	64	9	41
HMRZF	bino	19	16	84	12	63	9	47
	diverg	19	13	68	9	47	9	47
	disty	19	12	63	7	37	6	32
	pid	19	15	79	11	58	9	47
HMRCZF	Bino	11	9	82	7	64	6	55
	Diverg	11	9	82	7	64	3	27
	Disty	11	9	82	8	73	3	27
	Pid	11	9	82	7	64	3	27

Table 5.3: Accuracy of detection of regulatory regions. The number of genes tested is shown for each set, together with the number and percentage of genes for which the regulatory region were detected. 'nrg' : number of human genes in the set ; 'nrpg' number of genes for which the regulatory region was detected (i. e. 'positive genes') ; '%pg' : percentage of genes for which the regulatory region was detected. Names of sets and scores are as in Table 5.2.

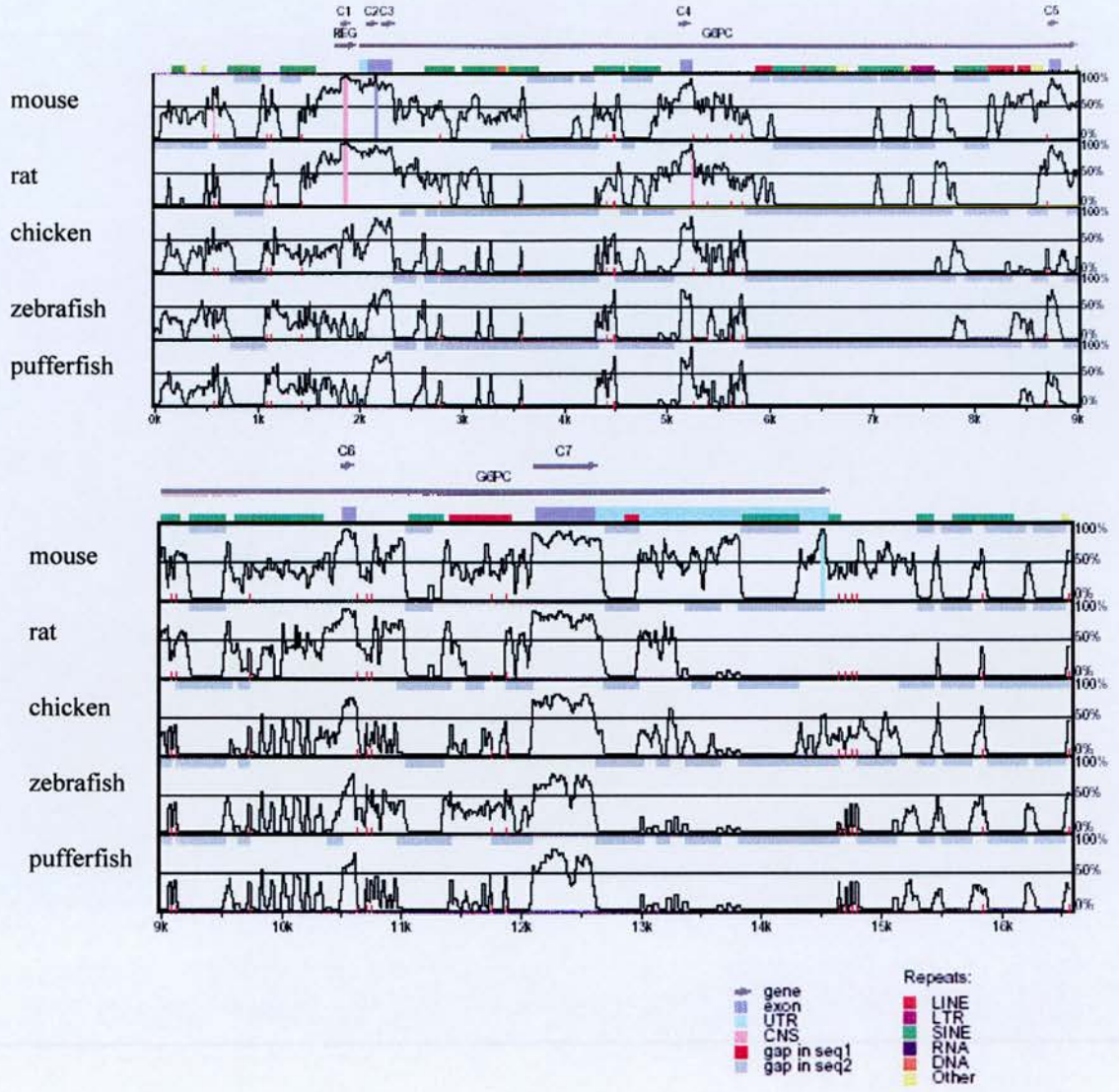


Figure 5.3: Percentage of identity plot of the genomic sequences of the human *G6PC* and vertebrate orthologs. The picture shows the multiple alignment of the *G6PC* gene sequences from human, mouse, rat, chicken, zebrafish and Fugu. Each pairwise alignment was derived from the multiple alignment of the sequences from the six species. Arrows on top on the graph indicate known genomic features on the sequence (*G6PC* gene, arrow 'G6PC', and the regulatory region, arrow 'REG') as well as the seven regions selected based on their level of conservation as measured by the probability of observing the calculated percentage of identity given the estimate of the rate of neutral evolution ('C' stands for conserved sequence).

5.4 Sequence conservation in dog, opossum and frog

To assess the reliability of the predictions, the sequences derived from the analysis of the HMR set were searched against the recently available draft genomic sequence of three vertebrates: *Canis familiaris* (dog), *Monodelphis domestica* (opossum), and *Xenopus tropicalis* (frog). As the available genomic sequence data for these species was whole genome shotgun reads, human repeat-masked sequences were searched against them using BLASTN (Altschul et al., 1990). Five types of human sequences were considered: i) known functional sequences, such as coding sequences and regulatory regions, ii) known regulatory regions alone, iii) putative novel functional sequences identified here ('novel' hereafter), iv) sequences with no known function and which were not selected by the HMR set analysis ('unknown' hereafter), v) randomly shuffled versions of the unknown sequences (with the exception of masked repeats, 'shuffled' hereafter). Only sequences larger than 40 bp were considered. Table 5.4 shows that only a single conserved sub-sequence was found in the shuffled sequences. Known functional sequences were most frequently conserved and a greater proportion of their length was conserved compared to the other sequences. The percentage of novel sequences that were conserved was lower than that for the known regulatory regions. However, the proportion of the length of sequence conserved was similar. Thirty two genes out of the 40 control genes (80%) contained known regulatory sequences with sub-sequences conserved in at least one species and 19 out of the 36 genes (53%) with novel sequences contained 'novel' sequences conserved in at least one species. Over these 19 genes, on average, 52% of the total length of the 'novel' sequences was conserved. Also, 95% of the genes contained 'unknown' sequences conserved in at least one species. Figure 5.4 shows that the quality of the conservation of novel sequences was higher than that of known functional sequences, in turn similar though slightly higher than that of 'unknown' sequences. This emphasises that functional regions may remain undiscovered even in well studied genes.

	regulatory			functional			novel			unknown			shuffled		
	count	perc_nb	perc_lgth	count	perc_nb	perc_lgth	count	perc_nb	perc_lgth	count	perc_nb	perc_lgth	count	perc_nb	perc_lgth
opossum	15	95.8	33.5	292	66.6	46.5	20	30.5	31.3	42	9.9	0.6	1	1.4	0
boxer	43	96.2	50.0	391	90.1	74.5	76	49.1	51.6	577	34.6	5.4	1	1.8	0
frog	2	100.0	23.3	59	43.8	27.0	4	12.9	18.4	5	3.1	0.2	1	1.4	0
≥ 1 species	44	96.3	79.3	400	92.2	76.3	76	49.1	51.2	588	35.0	5.5	1	2.3	0
≥ 2 species	12	95.5	56.5	283	66.9	61.1	20	30.5	35.8	28	9.1	1.4	0	0	0
3 species	2	100.0	54.5	146	42.7	42.5	4	12.9	21.5	2	3.1	0.6	0	0	0

Table 5.4: Conservation of human genomic sequences in dog, opossum and frog. 'count': number of conserved sub-sequences; 'perc_nb': mean of the percentage of query sequences with at least one conserved sub-sequence, over genes with at least one conserved sub-sequence in at least one query sequence for that type of annotation feature; 'perc_lgth': mean of the proportion of the length of the query sequences represented in the conserved sub-sequences, over genes with at least one conserved sub-sequence in at least one query sequence for that type of annotation feature.

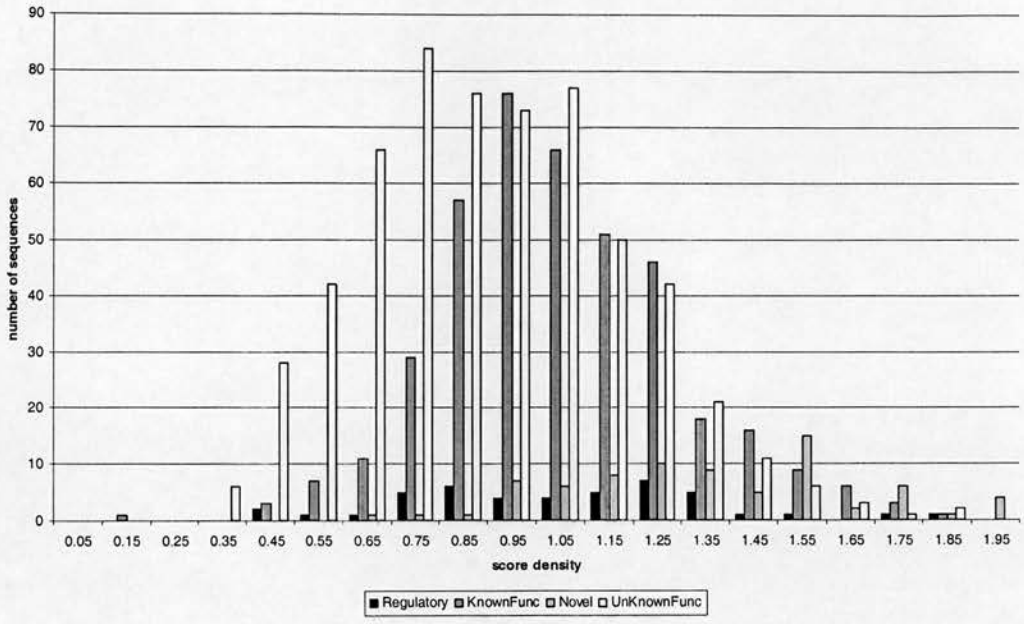


Figure 5.4: Score density of the sub-sequences conserved in dog, opossum, or frog. Distribution of the score density of sub-sequences conserved in at least one of the following species: dog, opossum and frog (score density = score / length). 'Regulatory': sequences with a known regulatory function; 'KnownFunc': sequences with a known function; 'Novel': sequences with no known function and selected by the analysis of the HMR set using the 'binomial' scoring system; 'UnKnownFunc': sequences with no known function and not selected by the analysis of the HMR set using the 'binomial' scoring system.

5.5 Discussion

Simple scoring systems based on weighting the average percentage of identity across a window scanning the multiple alignments showed a modest sensitivity and high specificity in detecting coding and regulatory regions. The scoring system based on the binomial probability of the amount of the observed conservation, given the rate of synonymous substitution (Thomas et al., 2003), was more accurate than the simpler scoring schemes in the conditions used here, as it was at least as sensitive as and as specific as the simpler scoring schemes. The computation of the binomial score partly relies on the estimation of dS. The use of more elaborate models of molecular evolution may however bring little improvement, as more complex models generally harbour higher variances than simpler, less realistic models (Nei and Kumar, 2000). In addition, an accurate estimation of dS appears not to be crucial as the sets of orthologs contained sequences from chicken and fish, which are saturated with synonymous substitutions. In any case, just as Margulies *et al.* (Margulies et al., 2003) demonstrated the value of a scoring system based on the rate of neutral evolution and binomial probability with a spectrum of mammalian species, this work shows here that a similar scoring system can perform well with a set of few distantly related species. The spread of the distribution of the average percentage of identity of the coding sequence in the sets containing chicken and fish sequences is wider than that in the HMR set, reflecting the fact that the use of distantly related species can blur the distinction between the non-coding and coding sequences in terms of average percentage of identity (Thomas et al., 2003). The inclusion of the chicken and fish sequences in the analysis reduced the number of genes analysed and had a slight detrimental effect on the accuracy of detection of regulatory regions, for all scoring methods. The high proportion of genes for which the regulatory region was identified using the chicken and fish sequences is likely due to the fact that where orthologs of the human gene could be identified in these species, the relevant genes where highly conserved.

Accuracy measured in terms of bases was on average lower with windows of 20 bp than with windows of 40 bp, while regulatory regions were detected in up to four more genes with the small windows compared to the 40 bp windows. This might be explained by the fact that gene transcription is mainly regulated by modules of

transcription factor binding sites, but not individual sites themselves (Yuh et al., 1998). The 40 bp window size may thus provide a better indication of functional conservation. The number of genes for which regulatory regions were detected using a 100 bp window was half that for shorter windows. This may suggest that conserved regions are shorter than 100 bp. While specificity is high for all genes, sensitivity varies widely among genes, reflecting the lack of conservation of all regulatory regions across all species for some genes (Yuh et al., 1998; Dermitzakis and Clark, 2002), and indicating that the approach is more appropriate and accurate for some genes than for others. Overall, the use of a threshold based on the proportion of coding region in the reference sequence analysed and on the proportion of the length of the human sequence selected (see Chapter 2) generated very high specificity and modest sensitivity, in spite of being conservative. The analysis of the mouse genome suggested that coding sequences represent only half of the functional genomic sequences (Waterston et al., 2002). Sensitivity might therefore be increased with a less stringent selection of conserved windows by using a threshold twice that of the proportion of coding regions in the human sequence. Such a threshold may however not be appropriate as it implies the assumption that functional sequences are evenly distributed and conserved across the genome, whereas the amount of sequence under selective constraint varies across the genome (Waterston et al., 2002).

As the set of known regulatory sequences available to date is limited, it may not be representative of all regulatory regions in the genome. This limitation is inherent to the current state of the knowledge and should diminish progressively with increased experimental verification of putative regulatory region function. It is noteworthy that these scoring schemes generated reliable results despite the necessarily imperfect annotation of gene structures provided automatically by Ensembl.

The reliability of these predictions was further assessed by studying the sequence conservation in other vertebrates where a considerable amount of genomic sequence is available. Novel sequences selected in the analysis which have no known function are more likely to be functional if conserved in these additional species. Results showed that 49 % of such sequences are conserved in at least one additional species and 30 % are conserved in at least two. Depending on the species involved,

these sequences may be either functional or part of the core neutral eutherian genome (Rat Genome Sequencing Project Consortium, 2004). Based on the data from opossum, at least 30 % of these sequences are likely to be functional. Finally, many conserved sub-sequences in 'unknown' sequences would probably be classified as 'novel' with a less stringent threshold on the total length of sequence to be selected.

The protocol tested here allowed for predictions of the functional potential of non-coding DNA to be sufficiently reliable to suggest a limited number of putative functional regions for experimental verification with an acceptable and conservative measure of the false positive rate. This approach was therefore applied to genes in the 4p region linked to psychosis to identify a limited number of highly conserved, putatively functional regions.

CHAPTER 6

COMPARATIVE ANALYSIS OF GENES IN A CANDIDATE REGION FOR PSYCHOSIS

6 COMPARATIVE ANALYSIS OF GENES IN A CANDIDATE REGION FOR PSYCHOSIS

6.1 Introduction

Bipolar Affective Disorder (BPAD), also named manic depression, and schizophrenia are two major psychiatric illnesses, each affecting 1% of the population during their lifetime (Weissman et al., 1988). The molecular and cellular bases of these diseases are not well understood, due to their heterogeneity, lack of animal model, and general incomplete knowledge of normal brain function (Owen and Cardno, 1999). However, genetic susceptibility is an important factor, as indicated by family, twin and adoption studies (Owen and Cardno, 1999). A region on chromosome 4p was identified by our laboratory by a genome-wide linkage analysis of a large pedigree, F22, affected by affective disorders such as BPAD and recurrent unipolar depression (UP) (Blackwood et al., 1996). This finding was subsequently confirmed by other groups that also reported linkage to schizo-affective disorder and schizophrenia (Asherson et al., 1998). The group followed this work by constructing a contig encompassing 6.9 Mb of the 4p15-16 region (Evans et al., 2001; Le Hellard et al., 2001), including an 11 cM D4S394-D4S403 (5.8 Mb) critical region for BPAD and a single remaining ~300 Kb gap. The region was further characterised by a computational study of transcripts mapped to the contig (Semple et al., 2000; Evans et al., 2001). Access to three other linked families: F59, CF50 and F48 was obtained through collaboration (Williams et al., 1999; Detera-Wadleigh et al., 1999). Haplotype analysis performed by the group in these four families allowed the determination of the haplotype that segregates with the disease in each of them.

The largest candidate region is bound by the recombination breakpoints in F22 and recombination breakpoints in the other families delineate four sub-regions. Two of these regions can be prioritised for candidate gene analysis by the strength of evidence provided by the different families. Figure 6.1 represents the 4p15-16 candidate region and shows the location of the two high-priority sub-regions referred to as 'region B' and 'region D' hereafter. The key activity of the group now consists

in carrying out case-control association studies of single markers and haplotypes in large control, BPAD and schizophrenia samples. SNPs used in these studies are either retrieved from public resources, such as dbSNP (Smigielski et al., 2000), or detected in-house in tested samples. SNPs that have been analysed so far were mainly located in the coding sequences of the genes in these two regions. However, evidence for the crucial role of regulatory polymorphisms in normal function of the cell and, and hence diseases, is now increasing (Wray et al., 2003). Identification of the causal variants in the 4p region therefore also relies upon the testing of SNPs in non-coding sequences. The large size of the candidate region, which spans approximately 8 Mb in total for both high-priority regions, underlines the necessity of narrowing the number of markers to test to a manageable number. Detection of non-coding sequences with a high potential for function is therefore indispensable to association studies. The work presented in this chapter aimed to apply the most successful protocol described in the previous chapter to identify a limited number of non-coding sequences to be screened for SNPs.

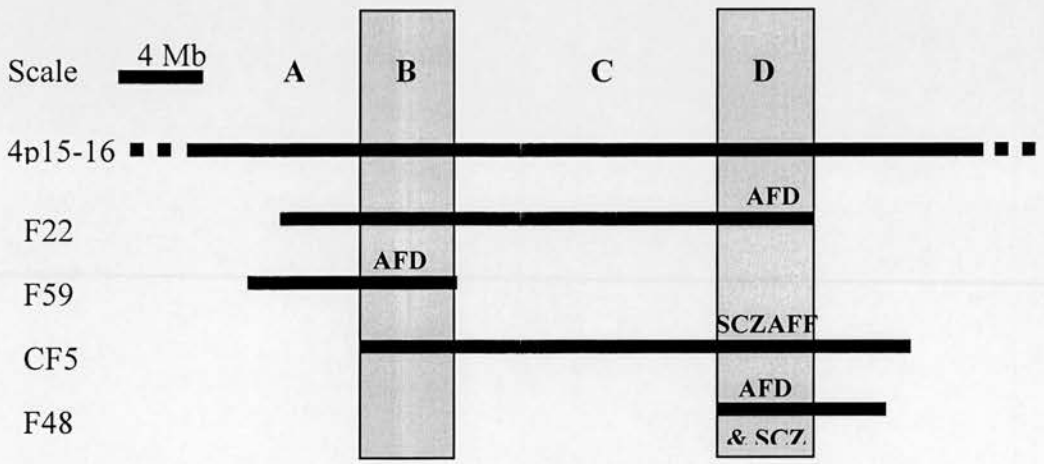


Figure 6.1: Candidate region for psychosis on chromosome 4p. The schema represents the 4p15-16 region linked to psychosis, together with the extent of the haplotypes that segregate with the diseases in the four families. The overlaps of these haplotypes and the strength of evidence define two high-priority sub regions D and B (vertical boxes), then C and A. 'AFD': major affective disorder; 'SCZAFF': schizo-affective disorder and schizophrenia; 'SCZ': schizophrenia.

The analysis reported in this chapter focused on regions B and D and specifically on the known genes and well-supported predicted genes. In human, regions B and D contained 9 and 17 such genes, respectively, as provided by Ensembl in April 2004. The data set analysed comprised all 26 human genes. For each of these genes, the orthologs in mouse, rat, chicken, zebrafish and the pufferfish *T. rubripes* were primarily identified using resources from Ensembl (similarity relationship, online BLAST facility) (Clamp et al., 2003), but also from the NCBI (including Homologene, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) and the UCSC Genome Bioinformatics site (Kent et al., 2002). Inspection of the conservation of synteny in this region was carried out to resolve conflicting data often due to the presence of paralogs and incorrect annotation. Genomic sequences of the genes identified and their 10 Kb flanking regions were retrieved from Ensembl. For each human gene, a multiple alignment of its sequence and those of its orthologs was built using MLAGAN (Brudno et al., 2003). A 40 bp window for which the 'binomial' score was computed then scanned the alignment. Top-scoring windows were then selected and sequences with high potential for function were derived (see chapter 5).

6.2 Sequence conservation in rodents, chicken and fish

6.2.1 Identification of orthologs

Orthologs of human genes in mouse, rat, chicken, zebrafish and pufferfish were identified using the similarity relationships provided by Ensembl, based on best reciprocal hits. Many cases were observed where the orthologs could not be identified readily, due to conflicting data regarding the available sequence assembly or annotation and regarding the suggested orthology relationships. Although these problems are related to one another, cases observed were classified into three types: i) sequence annotation and assembly, ii) sequence availability, and iii) orthology relationships, each sub-divided into several categories, as shown in Table 6.1. Out of the seven mis-assemblies observed five where found in the mouse (assembly NCBI32), of which four were not present in the NCBI30 assembly of the mouse genome. Evidence for the incorrect prediction of first exons was observed for 16 genes, based on EST data and truncated open reading frames. Inspection of the conservation of synteny led to the exclusion of one ortholog reported in rat, eight

reported in zebrafish, and five reported in pufferfish. For six human genes, while no ortholog was reported, inspection of synteny of the flanking genes in the species of interest allowed the identification of the putative ortholog in that species. However, in seven additional cases, synteny of the flanking genes was conserved in the species but the ortholog could not be identified, due to the presence of a gap in the sequence, or because the gene is not present in the organism or at that position, or because the sequence was mis-assembled.

	Key	Description
Annotation and assembly	ann1	**Mis-assembly**: the predicted structure of the gene was spurious, based on the annotation and structure of close orthologs (mouse); or the sequence was fragmented (chicken).
	ann2	**Gene split**: two genes were predicted that spanned the most likely actual gene.
	ann3	**Missed 5' end**: The structure of the Ensembl gene differed from that of the NCBI gene, and/or ESTs and promoter predictions by FirstEF and Eponine suggested an alternative start of the gene at a position further upstream than that predicted. This derived gene structure was further supported by an ortholog of the gene.
Sequence data	seq1	**Truncated gene**: The sequence of the reported ortholog seems to be truncated and too short to be used in the analysis.
	seq2	**Truncated flanking sequence**: the reported ortholog was located in a sequence such that the flanking region available was shorter than 10 Kb.
Orthology	ort1	**Spurious ortholog**: conservation of synteny of the reported ortholog and its flanking gene(s) i) could not be assessed, due to missing sequence or lack of convincing evidence for homology to any other gene, ii) suggested that the reported genes was in fact the ortholog of a different human gene, or iii) provided no or weak support for the orthology relationship reported.
	ort2	**Too many orthologs**: several orthologs were reported, and synteny conservation was used to select the most likely true ortholog.
	Ort3	**Not reported but suggested**: no ortholog was reported but conservation of synteny of the genes flanking the human gene, in the species of interest suggested that the ortholog may exist but could not be found, because of a gap in the sequence, or a putative mis-assembly.
	Ort4	**Not reported but detected**: no ortholog was reported but conservation of synteny of the genes flanking the human gene, in the species of interest suggested the presence of the ortholog, which was further supported by expression data and gene prediction.

Table 6.1: Categories of problems observed in the identification of vertebrate orthologs of known human genes in the two chromosome 4 candidate regions. 'Key': names of categories used in Table 6.2; 'Description': description of the problem.

Type	Sub-type	Nr	Gene	Species
Annotation, assembly	mis-assembly	7	CPZ	mouse (NCBI30)
			SLC2A9	mouse
			MIST0	mouse (NCBI30)
			DKFZp761B107	mouse (NCBI30)
			KIAA0746	chicken
			RBPSUH	chicken
			FLJ11082	mouse (NCBI30)
	gene split	2	GPR125	rat
			ZCCHC4	human
	missed 5' end	16	KIAA1729	human, mouse, rat, chicken
			GBA3	rat
			HS3ST1	rat, chicken, zebrafish, pufferfish
			PPARGC1A	human, mouse
			DKFZp761B107	rat, chicken
			ZCCHC4	human
			SLC34A2	human
			KIAA0746	human
Sequence data	truncated gene	3	SLC2A9	pufferfish
			SLA/LP	zebrafish
			CCKAR	rat
	truncated flanking sequence	7	WDR1	rat
			GBA3	chicken, pufferfish
			HS3ST1	zebrafish
			PI4K2B	pufferfish
			KIAA0746	chicken, pufferfish
Orthology	spurious ortholog	14	DRD5	zebrafish
			SLC2A9	zebrafish
			WDR1	zebrafish, pufferfish
			KIAA1729	zebrafish, pufferfish
			GPR125	zebrafish, pufferfish
			LGI2	zebrafish
			PIK2B,	zebrafish
			ZCCHC4	zebrafish, pufferfish
			RBPSUH	rat
			CCKAR	pufferfish
	too many orthologs	6	GPR78	pufferfish
			CPZ	pufferfish
			SLC2A9	pufferfish
			SLA/LP	rat
			RBPSUH	zebrafish
			CCKAR	zebrafish
	not reported but suggested	7	FLJ35	zebrafish
			GPR78	mouse, rat
			KIAA0746	zebrafish
			LOC389203	chicken
			RBPSUH	rat, pufferfish
	not reported but detected	6	GBA3	mouse, chicken, zebrafish
			PI4K2B	pufferfish
			LOC389203	mouse
			CCKAR	pufferfish

Table 6.2: List of problems observed in the identification of vertebrate orthologs of known human genes in the two chromosome 4 candidate regions. 'Nr' number of cases observed; '(NCBI30)': The sequence used was retrieved from the NCBI30 assembly of the mouse genome.

After inspection of the availability and quality of the annotation of the sequences in the six species, all 26 known human genes originally considered were included in the analysis. The number and total length of exons in the human sequences are displayed in Table 6.3. The genes included in the analysis are listed for each species in Table 6.4.

		Set of genes	Gene average
Coding sequence	Number	279	11
	Length (Kb)	45.1	1.8
5'UTR	Number	34	1
	Length (Kb)	6.7	0.3
3'UTR	Number	38	2
	Length (Kb)	24.3	1.0
Introns	Number	263	11
	Length (Kb)	1 690.5	67.7
Flanking regions	Number	50	2
	Length (Kb)	515.7	20.7

Table 6.3: Genomic sequence features in well-characterised genes in the two candidate regions. Table showing the number and total length of coding sequences, 5'UTRs, coding sequences, 3'UTRs, introns and flanking regions, summed over all genes in the HMR set (column 'Set of genes'), and averaged over these genes (column 'Gene Average').

6.2.2 Conservation of synteny

Conservation of synteny of human genes in regions B and D was inspected in mouse, rat, chicken, zebrafish and pufferfish. Analysis of region B suggests the presence of important rearrangements, or errors in the assembly of this region in the mouse, rat and chicken genomes, as suggested by differences in the order of some genes, absence of some genes, and a large insertion within one gene (Figure 6.2). Evidence for conservation of synteny for four genes was found in zebrafish and pufferfish (Figure 6.2). Synteny of the nine genes studied in region D, separated in two blocks, was conserved in mouse, rat and chicken, but not in fish.

Human			Mouse		Rat		Chicken		Zebrafish		Fugu	
HUGO ID	Ensembl Gene ID	Note	Ensembl Gene ID	Note	Ensembl Gene ID	Note	Ensembl Gene ID	Note	Ensembl Gene ID	Note	Ensembl Gene ID	Note
LOC389203	ENSG00000180408		ENSMUSESTG00000019454	ort4	ENSRNOG00000024656			Ort3				
SLC2A9	ENSG00000109667			ann1	ENSRNOG00000005302		ENSGALG00000014971			ort1		seq1
WDR1	ENSG00000071127		ENSMUSG00000005103	seq2	ENSRNOG00000028498		ENSGALG00000014957			ort1		ort1
KIAA1729	ENSG00000178163	ann3	ENSMUSG00000046572	ann3	ENSRNOG00000028534	ann3	ENSGALG00000014954	Ann3		ort1		ort1
MIST	ENSG00000109684		ENSMUSG00000039315	ann1	ENSRNOG00000028680		ENSGALG00000014951					
ZCCHC4	ENSG00000168228	ann3, ann2	ENSMUSG00000029179		ENSRNOG00000004025		ENSGALG00000014379			ort1		ort1
RBPSUH	ENSG00000168214		ENSMUSG00000039191			ort1,3		Ann1	ENSDARG00000003398	ort2		ort3
GBA3	ENSG00000176201		ENSMUSG00000000000	ort4	ENSRNOG00000024634	ann3		ort4, seq2	ENSDARG00000000000	ort4	SINFRUG00000150713	seq2
GPR78	ENSG00000155269			ort3		ort3	ENSGALG00000015595		ENSDARG00000016891		SINFRUG00000149474	ort2
CCKAR	ENSG00000163394		ENSMUSG00000029193			seq2	ENSGALG00000014363		ENSDARG00000003554	ort2	SINFRUG00000125923	ort1,4
SLALP	ENSG00000109618		ENSMUSG00000029173			ort3	ENSGALG00000014386		ENSDARG00000022830	seq1	SINFRUG00000133204	
FLJ35725	ENSG00000155275		ENSMUSG00000029097		ENSRNOG00000008898		ENSGALG00000015594			ort3	SINFRUG00000149473	
DRD5	ENSG00000169676		ENSMUSG00000039358		ENSRNOG00000005338		ENSGALG00000014975			ort1	SINFRUG00000131002	
LGI2	ENSG00000153012		ENSMUSG00000039252		ENSRNOG00000003887		ENSGALG00000014389			ort1	SINFRUG00000133206	
KIAA0746	ENSG00000091490		ENSMUSG00000029189		ENSRNOG00000004932	ann3	ENSGALG00000014370	Ann3, seq2		ort3	SINFRUG00000143177	
CPZ	ENSG00000109625		ENSMUSG00000036596	ann1	ENSRNOG00000008947		ENSGALG00000015597		ENSDARG00000015570		SINFRUG00000125509	ort2
HS3ST1	ENSG00000002587		ENSMUSG00000051022		ENSRNOG00000010598	seq3	ENSGALG00000014934	Seq3	ENSDARG00000018340	seq3, seq2	SINFRUG00000127547	seq3
GPR125	ENSG00000152990		ENSMUSG00000029090		ENSRNOG00000024538	ann2	ENSGALG00000014404		ENSDARG00000027931	ort1	SINFRUG00000144726	ort1
PPARGC1A	ENSG00000109819	ann3	ENSMUSG00000029167	ann3	ENSRNOG00000004473		ENSGALG00000014398		ENSDARG00000008721	(seq2)	SINFRUG00000150714	(seq2)
DHX15	ENSG00000109606		ENSMUSG00000029169		ENSRNOG00000003844		ENSGALG00000014395		ENSDARG00000015392		SINFRUG00000140493	
SOD3	ENSG00000109610		ENSMUSG00000050180		ENSRNOG00000003869		ENSGALG00000014393		ENSDARG00000006845		SINFRUG00000140486	
DKFZp761B107	ENSG00000181982		ENSMUSG00000051719	ann1	ENSRNOG00000024454	ann3	ENSGALG00000014391	Ann3	ENSDARG00000020857		SINFRUG00000140484	
PI4K2B	ENSG00000038210		ENSMUSG00000029186		ENSRNOG00000003924		ENSGALG00000014383		ENSDARG00000013881	ort1	SINFRUG00000150197	ort4 ,seq2
ANAPC4	ENSG00000053900		ENSMUSG00000029176		ENSRNOG00000004130		ENSGALG00000014378		ENSDARG00000012734		SINFRUG00000123629	
SLC34A2	ENSG00000157765	ann3	ENSMUSG00000029188		ENSRNOG00000004626		ENSGALG00000014372		ENSDARG00000012903		SINFRUG00000140865	
FLJ11082	ENSG00000109680		ENSMUSG00000039178	ann1	ENSRNOG00000003003		ENSGALG00000014359		ENSDARG00000013842		SINFRUG00000125927	

Table 6.4: Groups of vertebrate orthologs for well-characterised genes in the two candidate regions. Keys to 'notes' are in Table 6.1.

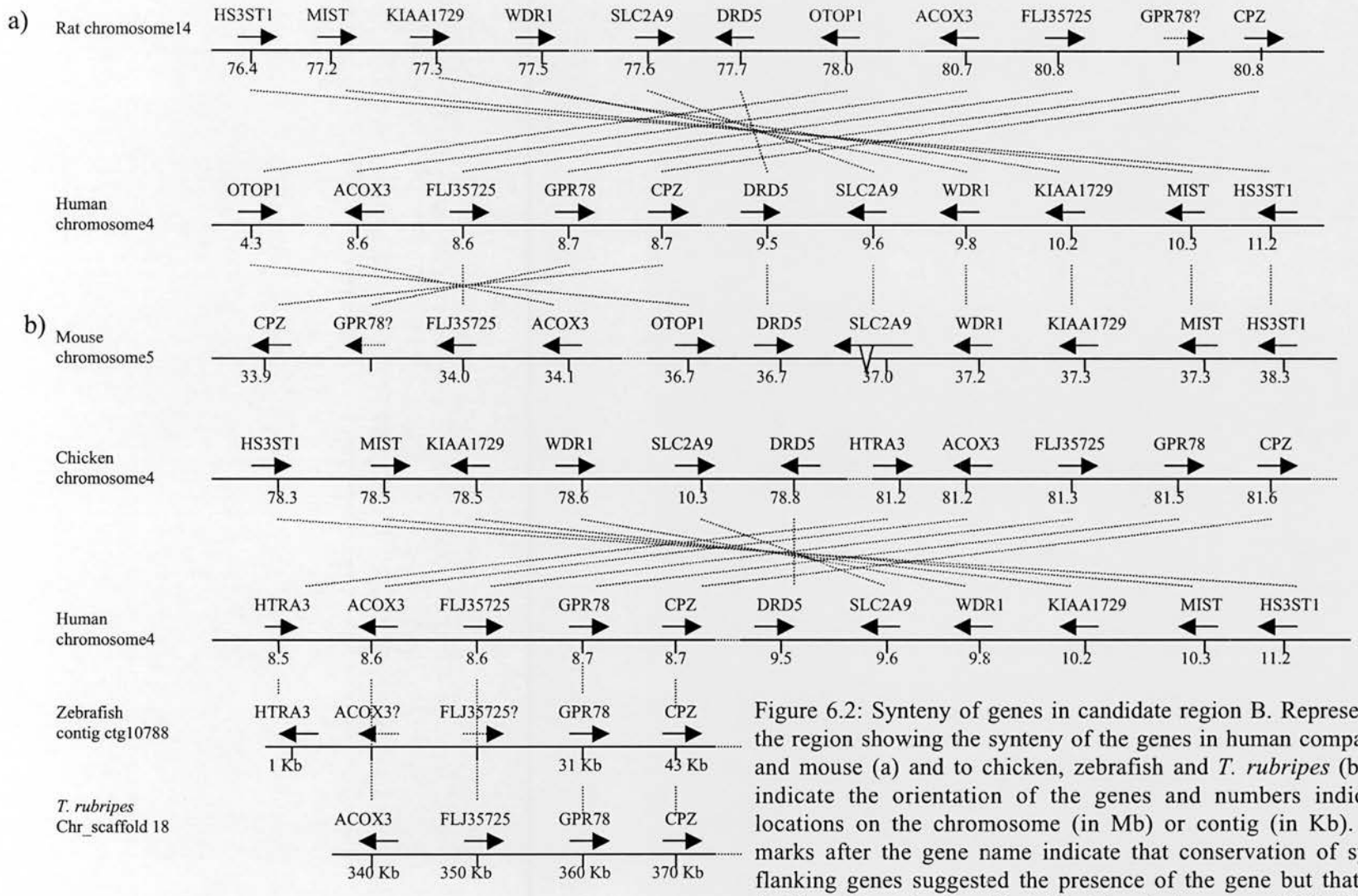


Figure 6.2: Synteny of genes in candidate region B. Representation of the region showing the synteny of the genes in human compared to rat and mouse (a) and to chicken, zebrafish and *T. rubripes* (b). Arrows indicate the orientation of the genes and numbers indicate their locations on the chromosome (in Mb) or contig (in Kb). Question marks after the gene name indicate that conservation of synteny of flanking genes suggested the presence of the gene but that the gene itself could not be identified. The triangle in mouse *SLC2A9* represents a 100 Kb gap and a 200 Kb insertion containing three predicted genes.

6.3 Highly conserved genomic sequences

Three sets of sequences were studied. First, the 'HMR' set consisted of each human gene and its orthologs in mouse and rat and contained 25 such groups of orthologs. Second, the 'HMRC' set consisted of the human genes and their orthologs in rodents and in chicken and contained 23 groups of orthologs. Third, the 'HMRCZF' set consisted of the human genes and their orthologs in rodents, chicken, zebrafish and Fugu. This set contains 20 sets of orthologs.

Results of the analysis are summarised in Table 6.5 and Table 6.6. The total length of human sequence analysed was 2.1 Mb, 1.9 Mb and 1.4 Mb for the HMR, HMRC and HMRCZF sets, respectively. The total length of sequence selected for a high potential for function represented 2.5, 2.6 and 3.0% of the sequence tested with the HMR, HMRC and HMRCZF sets, respectively. Over all genes, the total length of selected sequence was: 54.2 Kb, 47.7 Kb and 40.9 Kb for the HMR, HMRC and HMRCZF sets, respectively, in a total of 646, 534 and 469 conserved sub-sequences. On average, between 22 and 26 conserved sequences were identified in each gene (Table 6.5). Out of the 646 conserved sub-sequences selected based on comparison between human, mouse and rat, HMR set, 263 (40%) overlapped coding sequences. The proportion of conserved sub-sequences overlapping coding sequences was higher when chicken was included (HMRC set), 45% (240/534), or when fish sequences were considered (HMRCZF set): 49% (230/469). The proportion of base pairs in coding sequences overlapped by selected conserved sub-sequences was: 58%, 73% and 74% for the HMR, HMRC and HMRCZF sets respectively. With the HMR set, three conserved sub-sequences were identified in 5'UTRs and did not overlap coding sequences, whereas 24 conserved sub-sequences were found to be located in 3'UTR without overlapping coding sequences. This difference was also observed with the other sets (Table 6.5). As shown in Table 6.3, over all genes in the HMR set, 5'UTRs cover 6.7 Kb, whereas 3'UTRs cover 24.3 Kb. The 4-fold difference in the number of conserved sub-sequences in 3'UTRs compared to 5'UTRs may therefore be partly accounted for by the difference in length of these sequences, and hence maybe in the quality of the annotation of UTRs in the sequences, rather than by a higher sequence conservation in 3'UTRs than in 5'UTRs.

	HMR		HMRC		HMRCZF	
	Sum	Average	Sum	Average	Sum	Average
Total number	646	26	534	22	469	23
Coding sequences	263	11	240	10	230	1
Non-coding sequences	383	15	294	12	239	12
5'UTR	3	0	2	0	2	0
3'UTR	24	1	23	1	23	1
Introns	251	10	183	8	133	7
Flanking	77	3	73	3	70	3
UTRs; Introns; Flanking	28	1	13	1	11	1
Percentage Novel (%)	60		55		52	

Table 6.5: Repartition of selected genomic sequences conserved in vertebrates. Results are displayed for the sets HMR, HMRC and HMRCZF. For each set, values in the columns 'Sum' show results summed over genes in the set, and values in columns 'Average' show results averaged over genes in the set. 'Total number': total number of conserved sub-sequences selected; 'coding sequences': number of conserved sub-sequences (CSSs) in coding sequences; 'Non-coding sequences': number of CSSs not in coding sequences; 5'UTR, 3'UTR, 'Introns' and 'Flanking': number of CSSs entirely within each type of sequence; 'UTRs; Introns; Flanking': number of CSSs overlapping each type of non-coding sequences; 'Percentage novel (%)' : Percentage CSSs that did not overlap a coding sequence.

Out of all selected conserved sub-sequences in the HMR set, 58% (383/646) did not overlap any coding sequence. These sequences will therefore be referred to as 'conserved non-coding sequences' hereafter. The percentage of such sequences identified when chicken sequences are included decreased to 55% (294/534, HMRC set), and to 52% (239/469, HMRCZF set) when fish sequence were considered. The average number of conserved non-coding sequences identified per gene was 15, 12 and 12 for the HMR, HMRC and HMRCZF sets, respectively.

The number of conserved non-coding sequences observed within 10 Kb upstream of the first exon of genes was 43, 53 and 54 for the HMR, HMRC and HMRCZF sets, respectively. Out of these sequences, 23 (53%), 20 (38%) and 16 (30%) were located within 2 Kb immediately upstream of the start location of the genes.

	HMR		HMRC		HMRCZF	
	Sum	Average	Sum	Average	Sum	Average
Total ali. length (bp)	2 787 152	111 486	3 044 204	126 842	2 654 205	132 710
Total seq. length (bp)	2 104 892	84 196	1 905 385	79 391	1 369 254	68 463
Total pred. length (bp)	54 240	2 170	47 680	1 987	40 859	2 043
Percentage selected (%)	2.6	2.6	2.5	2.5	3.0	3.0
Coding bp selected (%)	58		73		74	
Non-coding bp in CSS (%)	52		42		38	
N coding bp selected	26 223	1049	27 862	1 160	25 335	1 268
N coding bp discarded	14 925	596	10 197	425	8 123	406
N non-coding bp selected	28 017	1121	19 818	825	15 524	776
N non-coding bp discarded	2 035 727	81 429	1 847 508	76 980	1 320 272	66 014

Table 6.6: Amount of selected genomic sequence conserved in vertebrates. Columns are as in Table 6.5. 'Total ali. length': total length of the aligned human sequence; 'Total seq. Length': total length of the original human sequence, 'Total pred. Length': total length of conserved sub-sequences; 'Percentage selected': percentage of base pairs in the human sequence selected as highly-conserved; 'Coding bp selected': percentage of coding base pairs selected; 'Non-coding bp in CSS': percentage of selected base pairs in non-coding sequences; 'N coding selected': number of base pairs in coding sequences selected; 'N coding discarded' : number of base pairs in coding sequences not selected, i.e. discarded; 'N non-coding selected': number of base pairs in non-coding sequences selected; 'N non-coding discarded': number of base pairs in non-coding sequences discarded'.

6.4 Polymorphisms in highly conserved genomic regions

Table 6.7 shows that 5 703 SNPs were identified in the genes in the HMR set, and 5 448 and 3 773 in the sets HMRC and HMRCZF, respectively. Out of these, 155, 123 and 107 were located in conserved sequences, representing approximately 2 to 3 % of the total number. These 'conserved SNPs' were classified according to whether the conserved regions overlapped a coding sequence or not. Within the HMR set, 41 out of the 155 conserved SNPs (26%) were located in conserved non-coding regions (defined as conserved regions that do not overlap a coding sequence).

	Single Nucleotide Polymorphisms			
	Total number	Conserved		
		Total Number	Coding	Non-coding
HMR	5703	155	114	41
HMRC	5448	123	80	43
HMRCZF	3573	107	76	31

Table 6.7: Number of single nucleotide polymorphisms in selected regions conserved in vertebrates. Names of sets are as in text.

6.5 Examples of highly conserved genomic regions

6.5.1 Anaphase promoting complex subunit 4, ANAPC4

Figure 6.3 illustrates results obtained for the analysis of the human '*anaphase promoting complex subunit 4*' (*ANAPC4*) gene in region D, which mediates metaphase-anaphase transition (Yu et al., 1998). Using sequences from six species, the 28 exons of this gene were all identified, by 29 CSSs (indicated by arrows heads in Figure 6.3). Out of the 12 remaining non-coding CSSs, 5 were located in introns, 2 in 3'UTRs, and 5 in the flanking regions. Three non-coding CSSs contained a SNP reported in dbSNP, and one also contained a putative binding site for the LIM Homeobox 3 (LXH3) transcription factor. Figure 6.4 shows the aligned sequences for these three CSSs and the information content at each position.

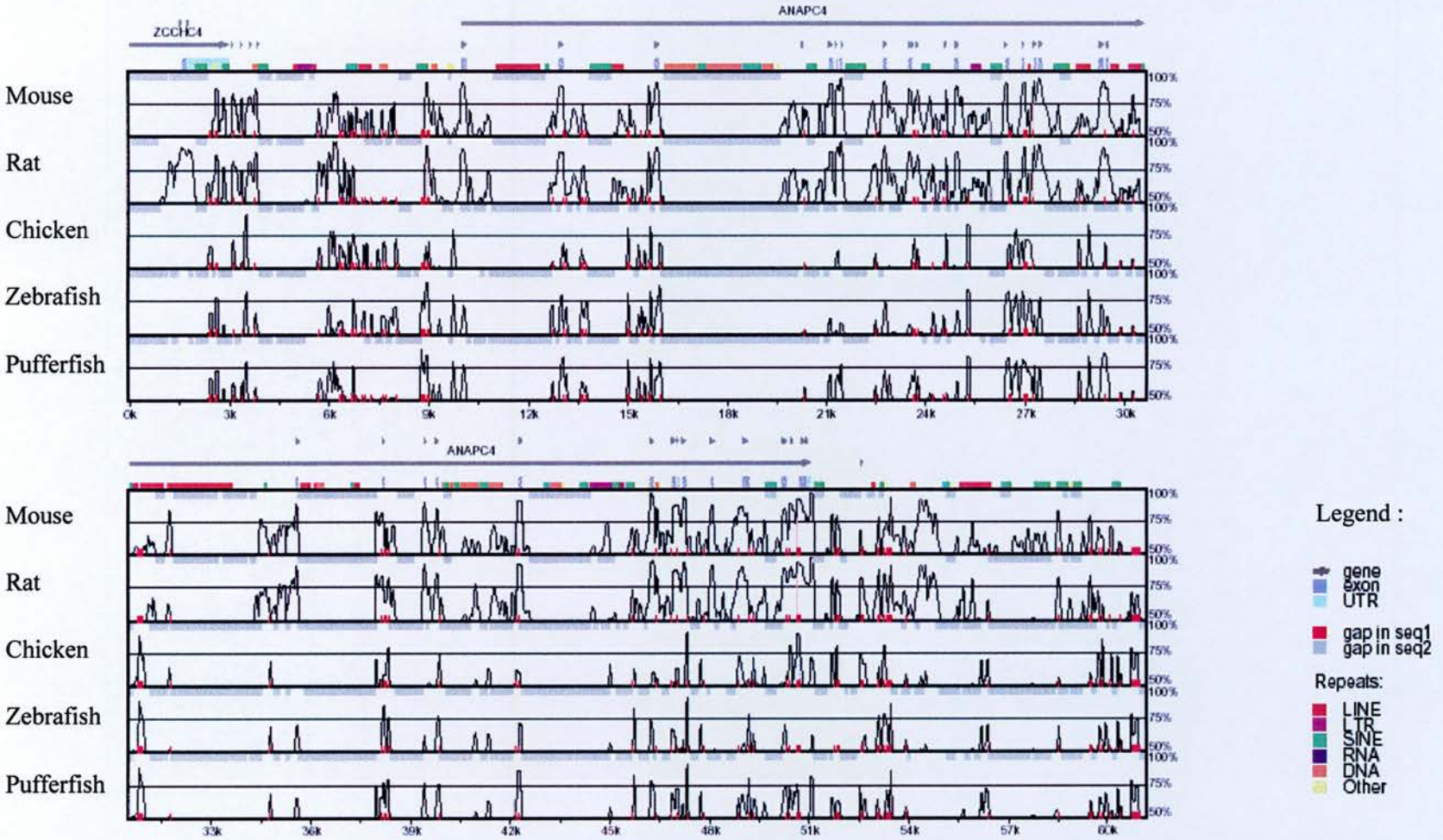


Figure 6.3: Percentage of identity plot of the genomic sequences of the human *ANAPC4* gene and vertebrate orthologs. The plot shows the percentage of identity of the sequence of the human *ANAPC4* gene and 10 Kb flanking regions with that of its orthologs in mouse, rat, chicken, zebrafish and pufferfish Fugu. Similarity was measured based on pairwise alignments derived from the multiple alignment of sequences from the six species. The large arrows indicate the location of genes (e.g. 'ANAPC4'), while small arrow heads represent conserved regions identified.

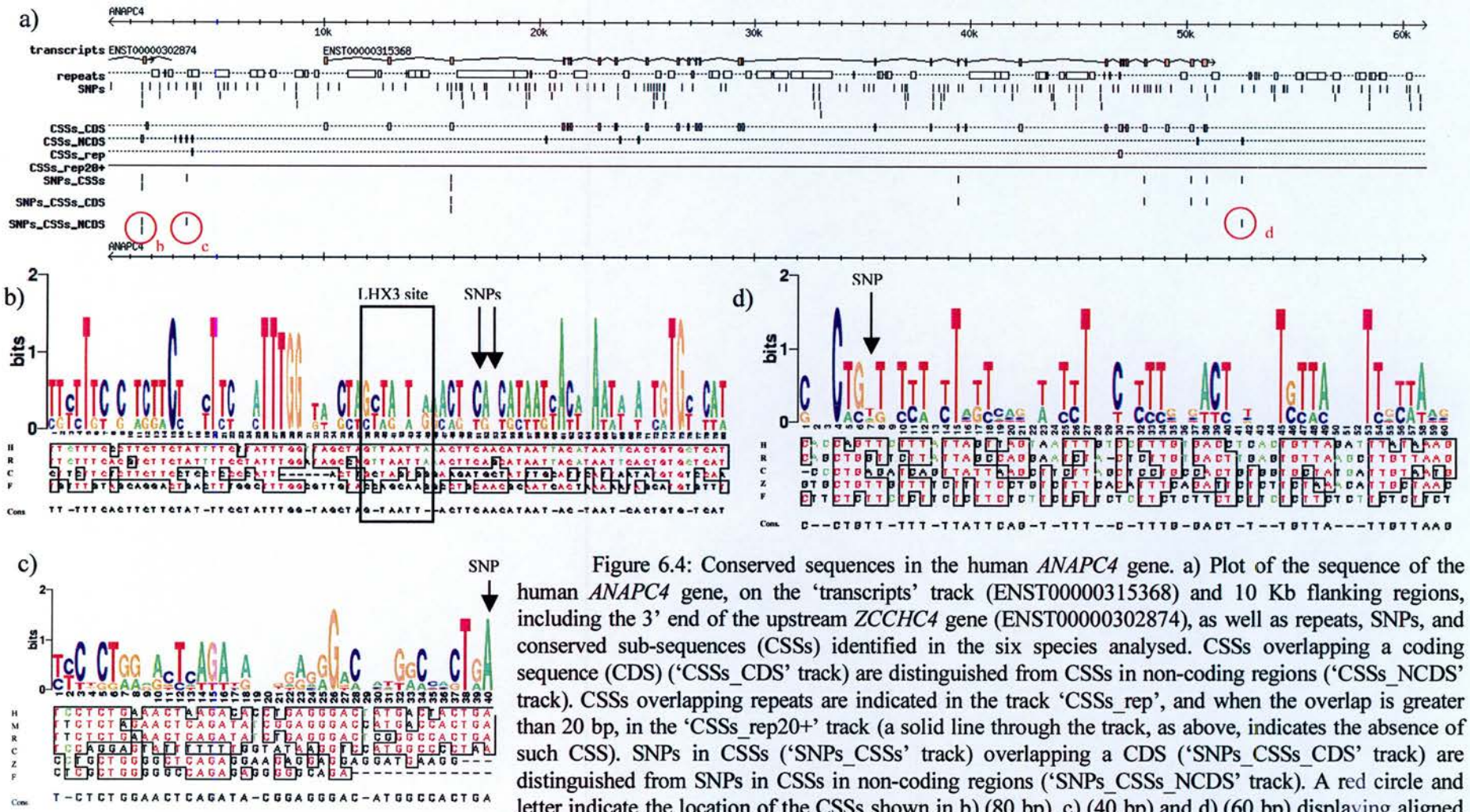


Figure 6.4: Conserved sequences in the human *ANAPC4* gene. a) Plot of the sequence of the human *ANAPC4* gene, on the 'transcripts' track (ENST00000315368) and 10 Kb flanking regions, including the 3' end of the upstream *ZCCHC4* gene (ENST00000302874), as well as repeats, SNPs, and conserved sub-sequences (CSSs) identified in the six species analysed. CSSs overlapping a coding sequence (CDS) ('CSSs_CDS' track) are distinguished from CSSs in non-coding regions ('CSSs_NCDS' track). CSSs overlapping repeats are indicated in the track 'CSSs_rep', and when the overlap is greater than 20 bp, in the 'CSSs_rep20+' track (a solid line through the track, as above, indicates the absence of such CSS). SNPs in CSSs ('SNPs_CSSs' track) overlapping a CDS ('SNPs_CSSs_CDS' track) are distinguished from SNPs in CSSs in non-coding regions ('SNPs_CSSs_NCDS' track). A red circle and letter indicate the location of the CSSs shown in b) (80 bp), c) (40 bp) and d) (60 bp) displaying aligned sequences (with identical bases boxed and red) in human (H), mouse (M), rat (R), chicken (C), zebrafish (Z) and Fugu (F), and the corresponding information content plot. 'Cons': consensus sequence; 'LHX3 site': LIM3 homeobox transcription factor binding motif.

Figure 6.4b shows the level of conservation observed for a CSS in the last intron of the *ZCCHC4* gene, at position 1.5 Kb in figure 6.4a. Figure 6.4b contains two parts: i) the aligned sequences, at the bottom, with identical letters in red and embedded in boxes, and ii) the corresponding plot of the information content, above the alignment, which illustrates the importance of individual sites in the alignment. The explanation of information content in this context is as follows. At a given column, one is uncertain of the bases in each sequence prior to reading them. While each base is read, the uncertainty decreases as some information is received. The measure of uncertainty is the base 2 logarithm of the number of symbols that can be found and is expressed in bits. If only one symbol is possible, uncertainty is zero, but with four equally possible symbols, such as nucleotides in a DNA sequence where gaps are discarded, maximum uncertainty is achieved as 2 bits. The information content is derived from the maximum uncertainty (i.e. before reading bases in the column) from which is subtracted the uncertainty after reading bases (i.e. accounting for the information received by reading bases) and a factor to account for the fact that a finite number of sequences have been used. In the ideal case of a high number of sequences and equiprobable nucleotides, a site with a base perfectly conserved has an information content of 2.

Although analysis of the *ANAPC4* gene included sequences from six species, human, mouse, rat, chicken, zebrafish and pufferfish, in the alignment shown in Figure 6.4b, sequences from mouse and zebrafish were discarded because they did not appear to contain the CSS. In Figure 6.4b, the T bases at positions 26 and 27 are conserved in the four sequences shown, and this is illustrated by a tall letter. Similarly, the neighbouring T and Gs at positions 28-30, are conserved in human, rat and zebrafish, but not chicken where these bases seem to be deleted. These positions are represented by a tall letter, since the sequence with the gap is ignored and only three sequences are considered, but the height of this letter is slightly smaller than for the T's present in all four species, reflecting the effect of the change in the number of sequences included. Similarly, position 25 contains three As but also a C, and is represented by much smaller letter A, reflecting the bigger effect of a substitution on the information content than that of a gap, as well as an even smaller letter C to indicate the substituted base. The information plot shown here is used to help read the alignment below it. This plot contains 80 positions in total, of which 9 harbour a tall letter representing the base conserved in four species, 41 smaller letters at positions where the represented base differs in one species. The region of the plot in

the box indicating the position of a putative LHX3, shows that out of the 8 positions in the motif, the 1st, 3rd, 4th and 6th columns have a base shared by three species. The total length of sequence shown for the 80 bp CSS (Figure 6.4b), 40 bp CSS (Figure 6.4c), and 60 bp CSS (Figure 6.4d, mouse sequence excluded) is 180 bp. Out of these 180 columns, 127 contained a nucleotide conserved in at least three species: 1 in all 6 species, 11 in 5 species, 46 in four species and 69 in three species. This illustrates the high conservation of these sequences, despite the presence of numerous, scattered substitutions in each sequence, shown in the seemingly unclear alignments but summarised in the information content plots. Region D contains two promising functional candidate genes: *cholecystokinin (CCK) A receptor* (*CCKAR*) and *phophatidylinositol 4-kinase type-II β* (*PI4K2B*).

6.5.2 Cholecystokinin (CCK) A receptor , CCKAR

The plot of the percentage of identity of the annotated sequence of the human *CCKAR* gene with the sequence of its ortholog in mouse, chicken, zebrafish and Fugu is shown in Figure 6.5. All five exons contained at least one CSS and several CSSs were detected in non-coding sequences: five in the 10 Kb upstream region, one in the 5'UTR, seven in introns, and two in the 10 Kb downstream region.

Details of four of the CSSs identified are shown in Figure 6.6: one CSS in a coding sequence (CDS) (Figure 6.6b), two in intron 2 (Figures 6.6c and 6.6d) and one in intron 3 (Figure 6.6e). In Figure 6.6b, the alignment of the 40 bp CSS in the second exon, at position 18 Kb in figure 6.6a, shows the level of conservation observed, which is illustrated in the corresponding plot of the information content, above the alignment. This plot contains 14 positions with a single tall letter representing the base conserved in all five species, 11 smaller letters at positions where the represented base differs in one species, and 6 even smaller letters at positions with 3 identical bases out of five. In comparison, the other three CSSs, in non-coding sequences, also harbour high levels of conservation with many bases conserved in all five species, such as in one of the 60 bp CSS in intron 2 (Figure 6.6c) located at 17.4 Kb in Figure 6.6a. Although, this sub-sequence does not seem to be present in Fugu, as indicated by gaps in the alignment in the bottom half of Figure 6.6c, it contains 9 bases identical in human, mouse, chicken and zebrafish, and 28 bases conserved in three of the four species.

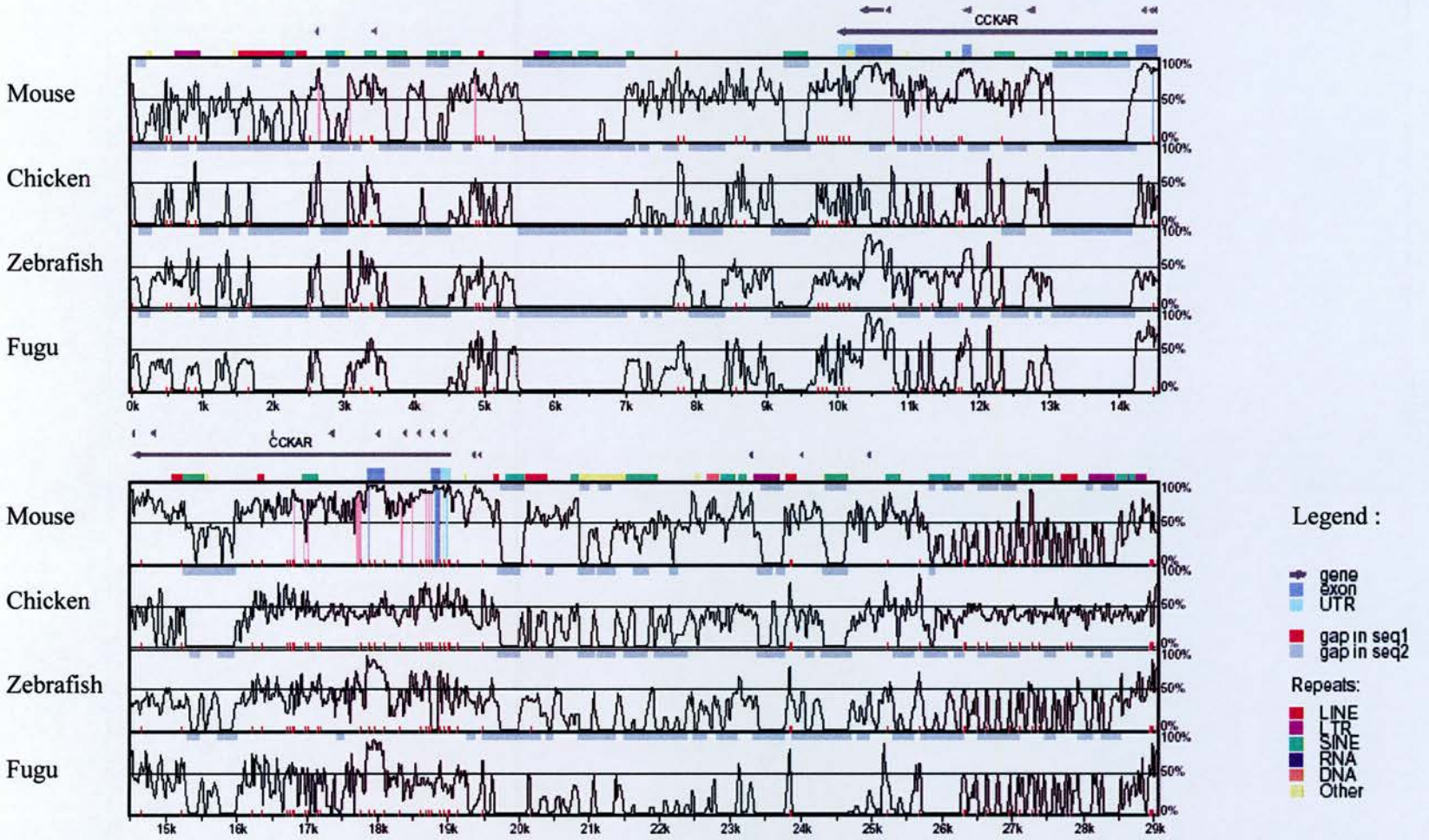


Figure 6.5: Percentage of identity plot of the genomic sequences of the human *CCKAR* gene and vertebrate orthologs. The plot shows the percentage of identity of the sequence of the human *CCKAR* gene and 10 Kb flanking regions with that of orthologs in mouse, chicken, zebrafish and pufferfish. Similarity was measured based on pairwise alignments derived from the multiple alignment of sequences from the five species. The large arrow indicates the location of the gene (CCKAR), while small arrow heads represent conserved regions identified.

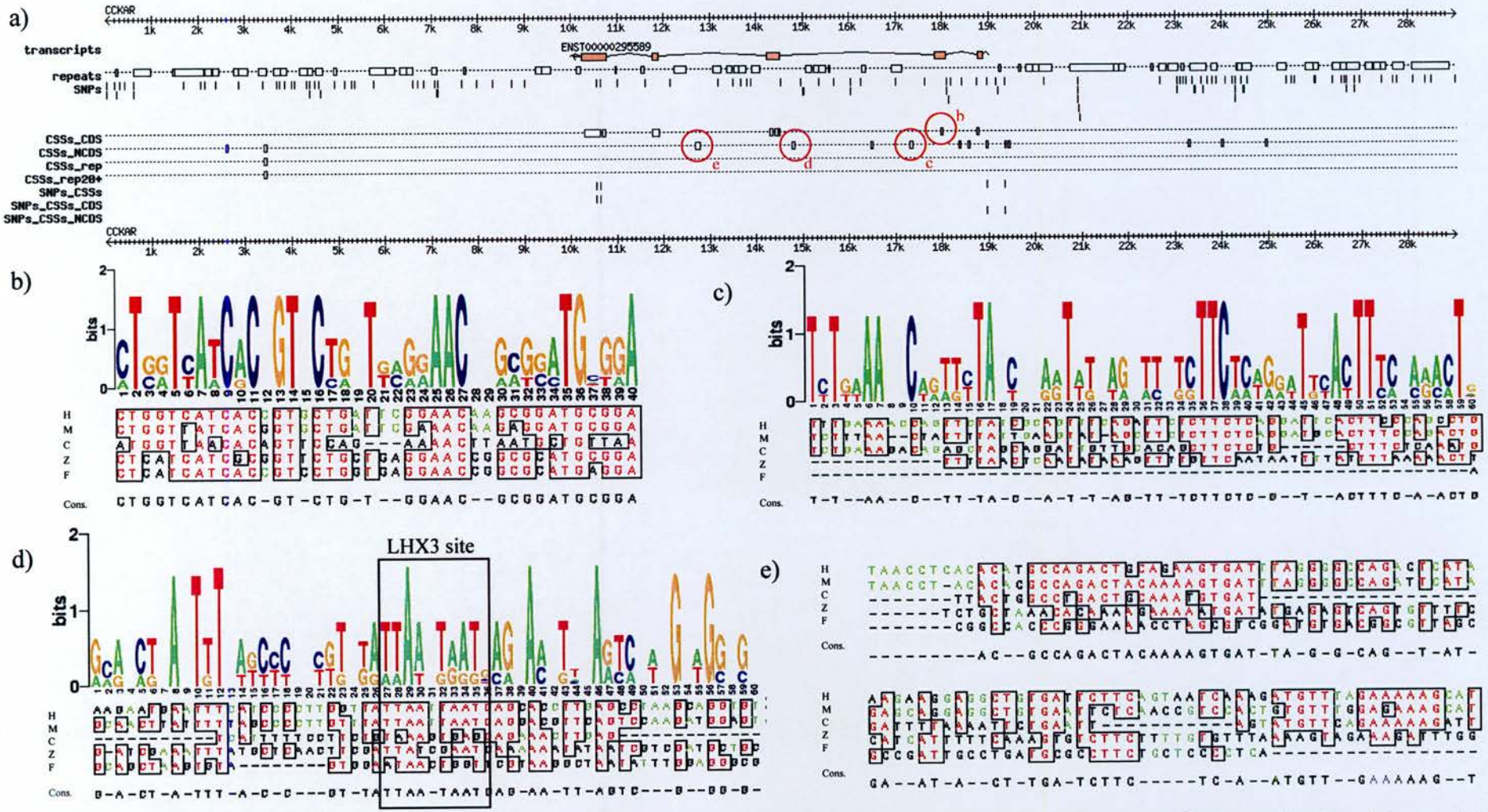


Figure 6.6: Conserved sequences in the human *CCKAR* gene. a) Plot of the sequence of the human *CCKAR* gene, on the 'transcripts' track (transcript ENST00000295589) and 10 Kb flanking regions, and conserved sub-sequences (CSSs) identified in mouse, chicken, zebrafish and pufferfish Fugu. Details as in Figure 6.4. b) 40 bp CSS in the second CDS; c) 60 bp CSS in intron 2; d) 100 bp CSS in intron 2, and e) 60 bp CSS in intron 3 (alignment only).

The alignment of the other 60 bp CSS in intron 2, at 14.7 Kb in Figure 6.6a and shown in Figure 6.6d, contains 38 columns with a base conserved in at least three species out of five: 4 columns with perfectly conserved bases, 15 columns identical in 4 species; and 19 columns with bases shared by 3 species. Gaps in one species are observed in 31 columns. This CSS also contains a putative binding site for the LIM Homeobox 3 (LHX3) transcription factor. Out of the 10 bases in this site, one is identical in all species and eight are conserved in four species.

The 100 bp CSS in intron 3, at 12.7 Kb in Figure 6.6a, is represented in Figure 6.6d. Ignoring the first seven bases present in human and mouse only, out of the 93 remaining positions, 45 contains a gap in chicken or Fugu. However, 60 of the 93 columns contain a base conserved in at least three species: one base is identical in all five species, while 26 columns contain a base conserved in 4 species and another 33 columns contain a nucleotide conserved in three species.

The upstream region of the *CCKAR* gene also harbours CSSs, including three that are located within 800 bp of the initiation codon. These three CSSs, 'CCS_4', 'CCS_5', 'CCS_6' and one CSS in the first intron, 'CCS_8' are shown in Figure 6.7a which displays the annotated sequence spanning 800 bp upstream and 800 bp downstream of the initiation codon, as well as the location of the CSSs. Alignments of these four 40 bp CSSs are shown in Figure 6.7b-e. Despite several gaps in these sequences, out of the 160 bp covered by the four CSSs, 99 bp are conserved in at least three species: 11 are identical in all five species, 23 are conserved in four species, and 65 are conserved in three species. Furthermore, out of the remaining 61 positions, 52 contain a nucleotide that is identical in human and mouse. Out of the 10 SNPs observed in this 1.6 Kb region, one lies in CSS_5, 'SNP_71' in Figure 6.7a, i.e. rs4349588 in Figure 6.7c, and another lies in CSS_6, 'SNP_68' in Figure 6.7a, i.e rs6448456 Figure 6.7d, at positions –473 and –85 from the initiation codon, respectively.

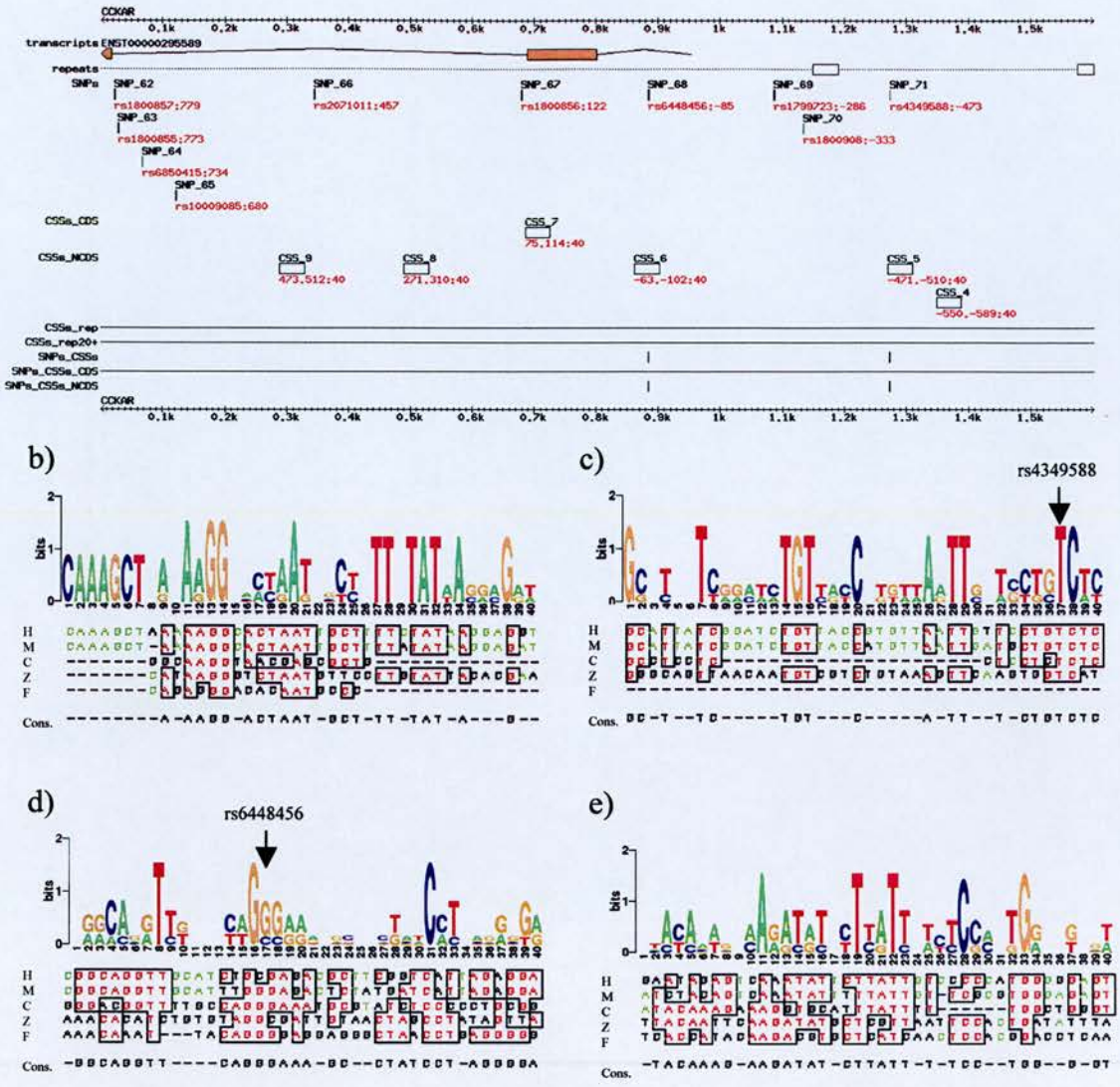


Figure 6.7: Conserved sequences in the 5' end of the human *CCKAR* gene. a) Plot of the region spanning 800 bp either side of the initiation codon of the human *CCKAR* gene, on the 'transcripts' track (transcript ENST00000295589), and conserved sub-sequences (CSSs) identified in mouse, chicken, zebrafish and pufferfish Fugu. SNPs are displayed with their internal name (black label), and their dbSNP name, followed by their position with respect to the initiation codon (red label). For example, at approximately 0.9 Kb on the scale, lies SNP_68, named 'rs6448456' and located at –85 with respect to the initiation codon (at 0.8 Kb on the scale). CSSs are labelled in black, and their location with respect to the initiation codon and length are shown in the red label (start.end:length). Remaining details as in Figure 6.4. b) 'CCS_4' in the upstream region, at 1.35 Kb on the scale; c) 'CSS_5' in the upstream region, at 1.25 Kb; d) 'CSS_6' in the 5'UTR, at 0.85 Kb; e) 'CSS_8' in intron 1, at 0.5 Kb on the scale.

6.5.3 Phophatidylinositol 4-kinase type-II β , PI4K2B

The percentage of identity plot obtained for the analysis of the *PI4K2B* gene in human, mouse, rat, chicken and Fugu is displayed in Figure 6.8. Out of the 28 CSSs identified, 12 were located in the 10 exons of the gene, 1 in the 5'UTR, 4 in the 3'UTR, 8 in introns, and three (11%) overlapped a repeat by more than 20 bp, two in the upstream region, and one in the downstream region. The latter three CSSs were each 40 bp and represented 6% of the total length of sequence covered by CSSs (1980 bp), while repeats represent 64% of the human sequence analysed (41 594 bp out of 65 061 bp). Across the 18 genes analysed in human, mouse, rat, chicken, zebrafish and pufferfish, 52 of the 471 CSSs (11%) overlapped a repeat, and for 41 of these, the overlap was longer than 20 bp. Half the CSSs overlapping a repeat by more than 20 bp were located in three genes: nine in *FLJ11082*, seven in *FLJ35725* and five in *GPR78*, for which repeats represent 56%, 48% and 24% of the human sequence analysed, respectively.

Details of four CSSs in non-coding regions of the *PI4K2B* gene sequence and 10 Kb flanking sequences are given in Figure 6.9, which also shows the entire annotated human *PI4K2B* gene sequence, and the position of CSSs (Figure 6.9a). The four examples shown are located on this sequence in intron 6 (Figure 6.9b), in intron 9 (Figure 6.9c) and in the 3'UTR (Figures 6.9d and 6.9e). The alignment of the CSS in intron 9 (Figure 6.9c) was manually edited by removing the sequence of the chicken because it did not contain the CSS and by adding a gap in the sequence of the zebrafish at position 20 and a gap in the other sequences at position 27. Out of the 181 bp covered by these four CSSs, 150 were conserved in at least three species: 32 were perfectly conserved, 58 were conserved in four species and 60 were conserved in three species. One of the CSSs in the 3'UTR contains two sub-sequences 'TTGTATTTTT', at positions 20-29 and 39-47 in the alignment shown in Figure 6.9e, both conserved in human, mouse and rat, with the first sub-sequence being conserved in chicken, too.

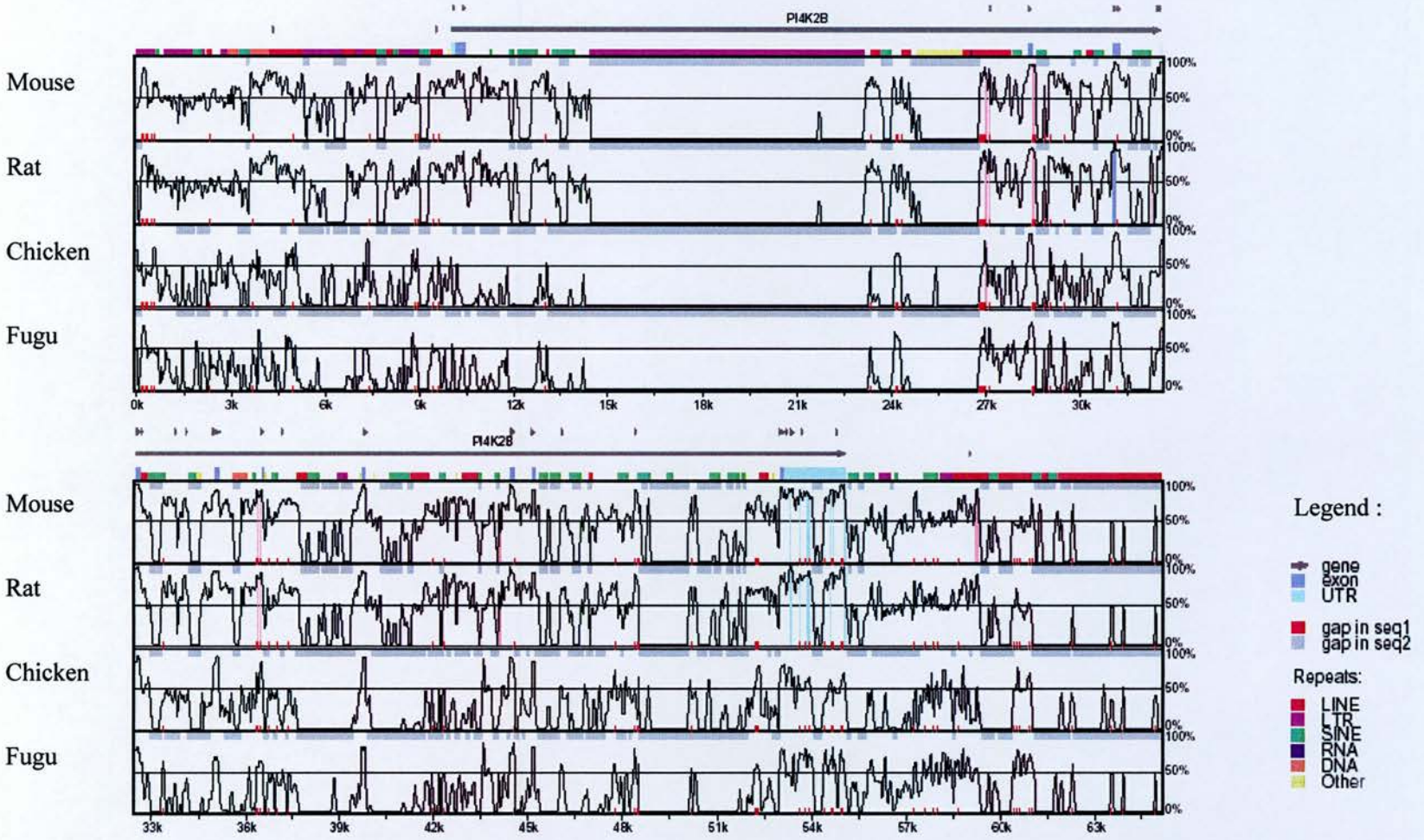


Figure 6.8: Percentage of identity plot of the genomic sequence of the human *PI4K2B* gene and vertebrate orthologs. The plot shows the percentage of identity of the sequence of human *PI4K2B* gene and 10 Kb flanking regions with that of mouse, rat, chicken and pufferfish. Similarity was measured based on pairwise alignments derived from the multiple alignment of sequences from the five species. The large arrow indicates the location of the gene (PI4K2B), while small arrow heads represent conserved regions identified.

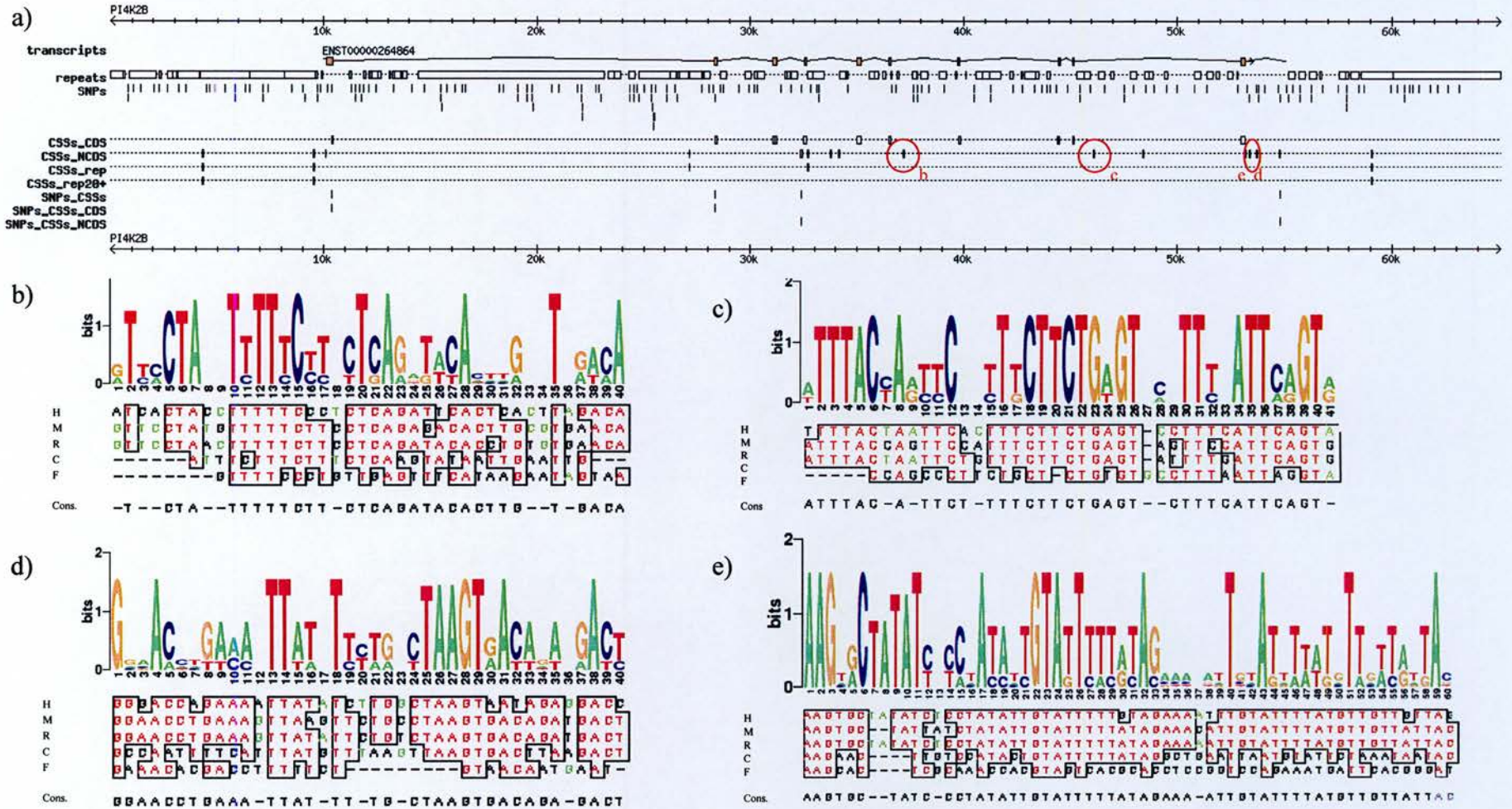


Figure 6.9: Conserved sequences in the human *PI4K2B* gene. a) Plot of the sequence of the human *PI4K2B* gene, on the 'transcripts' track (transcript ENST00000264864) and 10 kb flanking regions, and conserved sub-sequences (CSSs). Details as in Figure 6.4.

6.5.4 DEAH (Asp-Glu-Ala-His) box polypeptide 15 , DHX15

CSSs shown above show high conservation in several species, and CSSs in non-coding sequences are usually 40 bp to 80 bp. In contrast, some genes contain large non-coding CSSs, such as the *DEAH (Asp-Glu-Ala-His) box polypeptide 15* (*DHX15*) gene in region D, at 24.2 Mb on chromosome 4. As shown in Figure 6.10, the *DHX15* gene contains a 240 bp CSS in intron 4, immediately downstream of exon 4 (Figure 6.10b), and a 140 bp CSS in intron 9, immediately downstream from exon 9 (Figure 6.10c).

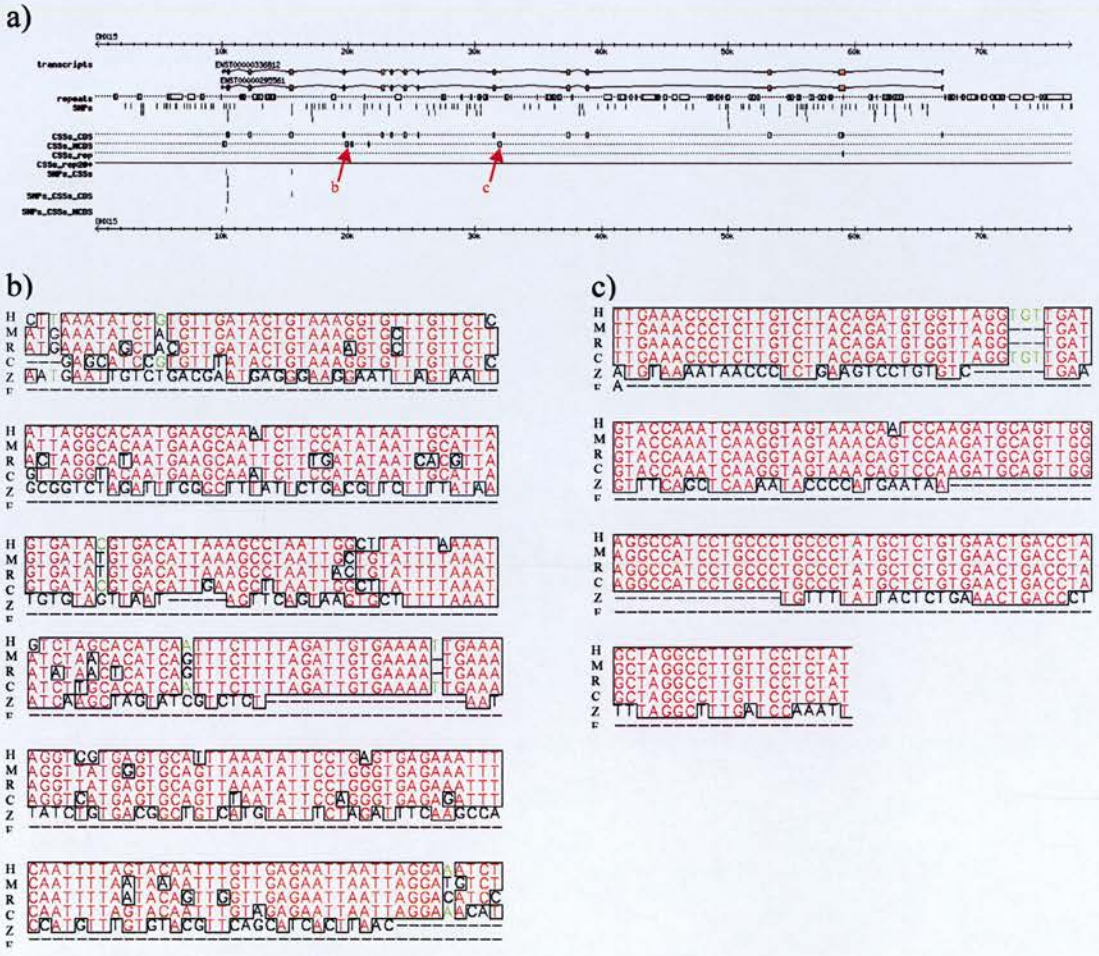


Figure 6.10: Conserved sequences in the human *DHX15* gene. Plot of the sequence of the human *DHX15* gene, on the 'transcripts' track and 10 kb flanking regions (a), and alignments of two intronic CSSs (b and c). Details as in Figure 6.4.

Alignments of the non-coding CSSs shown here for four genes appear to contain more gaps, and fewer positions where bases are conserved in all species than CSSs in coding sequences. Nonetheless, these four examples illustrate that these genes contain non-coding sequences of 40 bp to 240 bp that are conserved in vertebrates as distantly related to Mammals as bird and fish. Several CSSs contain SNPs and two of them also contain a putative binding site for the LIM Homeobox 3 (LXH3) transcription factor (Figure 6.4a and 6.6d).

6.6 Discussion

Comparative genomics has been reported to be an efficient tool to detect functional genomic sequences (Wasserman et al., 2000; Pennacchio and Rubin, 2001; Chapman et al., 2004). This chapter reports the identification of known and putative functional sequences in the human chromosomal region 4p15-16 linked to psychosis (Blackwood et al., 1996), by comparing the human sequence to sequences from five vertebrate species for which the genome sequence was publicly available. Results presented here were obtained by applying the protocol described in the previous chapter to genes with known function and genes with unknown function but supported by convincing evidence from gene prediction and expression data. This method aimed to detect sequences conserved to a greater extent than would be expected according to the rate of synonymous substitution measured for this gene. The protocol used can be briefly summarised as follows. First, orthologs of the human gene were sought in mouse, rat, chicken, zebrafish and pufferfish. Second, the protein sequences encoded by the human gene and by the orthologs identified in these species were aligned to estimate the rate of synonymous substitution at four-fold degenerate positions. A global multiple alignment of the genomic sequences for the gene and its flanking regions from each species was then built. Aligned sequences were subsequently scanned using a sliding window to identify regions showing higher conservation among the species considered than expected from the estimated rate of synonymous substitutions. Finally, SNPs in the conserved sequences detected were identified.

As for the genes used to benchmark the protocol described in the previous chapter, identification of orthologs of the human genes analysed here was cumbersome. Prior to the 'pre-release' of the chicken genome sequence by Ensembl, attempts were made to identify orthologs by developing a script to consult the Ensembl SQL database (version 15), analyse the reported best reciprocal hits and identify genes in mouse, rat, zebrafish and pufferfish that were each the best reciprocal hit in the other four species. This generated helpful information but manual inspection of the reported putative orthologs in each species remained necessary. In the easiest cases, problems were related to the structure of genes provided by Ensembl. Most discrepancies were resolved by inspecting annotation of the sequences in detail. This was possible for the genomes of the mouse, rat and chicken, although some orthologs, in rat and chicken in particular, were discarded due to errors in the fragmented genomic assemblies. Other problems related to the identification of the orthologs themselves, especially when no ortholog or several orthologs were suggested. Several problems of this type were resolved by searching the protein and genomic sequences from the different species with the homologous sequence in an attempt to identify the ortholog manually, and by inspecting synteny of the genes in all six species. Several problems could however not be deciphered, such as in the case of genes from large families, and in particular in the identification of orthologs in fish. In most cases, this most likely reflected the lack of sequence data, resulting in the fragmented assembly of these genomes, and also the presence of paralogs in these species.

Results obtained by comparing sequences from human, mouse, rat, chicken, zebrafish and pufferfish were overall similar to that obtained when fish sequences are not considered. However, fewer genes were included in the analysis using sequences from all six species as orthologs in fish could not be identified for six human genes. Conservation in fish of a sequence conserved in human, mouse, rat and chicken further supports the assumption that this sequence is functional. However, conservation of a sequence in rodent and chicken alone is strong evidence for putative function. Since the recent release of the assembly of the chicken genome, the labour required to identify orthologs in fish therefore seems to outweigh the additional knowledge provided by their analysis, at least until the quality of the

sequence data in these species improves greatly. Furthermore, genomes from several mammalian species, which are more closely related to human and therefore more appropriate for the analysis of human genes than chicken and fish, are being sequenced, such as the dog, of which the genome assembly is due to be published in the near future (http://www.nhgri.nih.gov). Sequences from these mammalian species will prove valuable as many functional genomic sequences required for biological processes specific to mammals, perhaps including brain functions involved in psychosis, may not be present in the genome of vertebrate species such as chicken or fish.

Several CSSs selected in this analysis overlap partially or entirely known interspersed repeats. It is possible that other CSSs overlap repeats not identified by RepeatMasker. However, RepeatMasker identifies known repeats reliably (http://www.repeatmasker.org) and it is therefore unlikely that known repeats exist in the sequences analysed but have not been detected. On the other hand, if some repeats were not identified, they would most likely belong to unknown classes of interspersed repeats. Such classes are expected to be rare since interspersed repeats in the human, mouse and rat genomes are well studied (International Human Genome Sequencing Consortium, 2001; Waterston et al., 2002; Gibbs et al., 2004). Furthermore, out of the 26 known human genes in the two candidate regions B and D, orthologous genes were identified in chicken and/or fish for 25 genes. Almost all CSSs identified are therefore shared between human and at least one distantly related species. The presence of interspersed repeats in these CSSs is unlikely because interspersed repeats are expected to accumulate mutations at a neutral rate (International Human Genome Sequencing Consortium, 2001; Waterston et al., 2002; Elnitski et al., 2003; Kolbe et al., 2004; Gibbs et al., 2004). As for CSSs detected here that overlap interspersed repeats, they may merely be due to the presence of sub-sequences within interspersed repeats that are conserved in vertebrates (Furano, 2000; Ogiwara et al., 2002). It is however noteworthy that certain classes of interspersed repeats are functional in eukaryotes such as Drosophila, sea urchin, mouse and human (Britten, 1997). These interspersed repeats may therefore be under selective constraint and presumably demonstrate higher conservation than expected. In primates, for instance, Alu repeats may be under

selective constraint related to their role in regulation of gene expression, rather than to the maintenance of their capacity to replicate in the genome (Britten, 1994).

Although most interspersed repeats are inactive in human, some are still active (Smit, 1996; International Human Genome Sequencing Consortium, 2001), and polymorphic insertions have been reported for several types of interspersed repeats (Carroll et al., 2001; Myers et al., 2002; Buzdin et al., 2003; Watkins et al., 2003). Some interspersed repeats insert within genes, disrupting their function and causing diseases (Druker and Whitelaw, 2004). For instance, insertion of a truncated element of the L1 subfamily of long interspersed nuclear elements (LINEs) into an exon of the *factor VIII* gene was detected in patients with haemophilia and proved to cause the disease (Kazazian et al., 1988). Interspersed repeats can also have a regulatory role when they insert in or near promoters and in enhancers of genes, as observed in human and mouse (Ostertag and Kazazian, Jr., 2001; Druker et al., 2004). Insertion of the 3'UTR segment of a L1 element in an enhancer 15 Kb upstream of the *growth hormone* gene cluster resulted in the expression of these genes (Shewchuk et al., 2001). Similarly, insertion of a L1 element in an enhancer situated 20 Kb upstream of the promoter of the *apolipoprotein A* gene, was shown to be responsible for a 10-fold increase in transcriptional activity of the gene (Yang et al., 1998). This appears to be due to the presence binding sites for the Sp1 and Ets transcription factors in the 5'UTR of L1 repeats (Yang et al., 1998). In addition, the 5'UTR of L1 elements also contains an antisense promoter that can drive transcription in the opposite direction from that of the internal L1, resulting in the production of transcripts with the adjacent flanking region (Speek, 2001). Because the human genome contains approximately 4000-5000 full-lengths L1 copies, it is possible that many of these elements influence the expression of genes nearby (Nigumann et al., 2002; Druker and Whitelaw, 2004). It is noteworthy that chromosome 4 appears to be much more permissive to insertions of the youngest type of L1 elements in human, specifically, than any other chromosomes, and than would be expected from its length or gene density, although the significance of this findings is not clear (Boissinot et al., 2004).

Short interspersed nuclear elements (SINEs) also contain a promoter region but no protein-coding sequence (Okada, 1991; Druker and Whitelaw, 2004). The

most prominent form of SINEs comprises Alu elements, which are the only SINEs still active in the human genome (International Human Genome Sequencing Consortium, 2001). Increasing evidence suggest that Alu repeats are implicated in transcription, translation, response to stress, nucleosome positioning and imprinting, and might have be involved in the evolution of regulatory networks in the primate lineage (Britten, 1994; Shankar et al., 2004). Several Alu insertions are thought to cause human diseases, including insertions occurring in coding sequences or close to exon boundaries, where they may affect splicing (Deininger and Batzer, 1999; Deininger and Batzer, 2002). For instance, an alternatively splice variant of *neuregulin 1*, one of the strong candidate genes for schizophrenia, contain an in-frame Alu-sequence in the coding sequence, but the biological relevance of this insertion is not known (Stefansson et al., 2004). Alu repeats also affect gene expression at the level of transcription, via insertions in or near promoter and enhancer regions (Britten, 1996), such as in an enhancer in the last intron of *CD8A* gene encoding the CD8 antigen, alpha polypeptide (p32) (Hambor et al., 1993), or in the promoter of the gene encoding the nicotinic acetylcholine receptor α6 subunit (CHRNA6) (Ebihara et al., 2002). Antisense transcription from the promoter of SINEs has also been observed in mouse (Ferrigno et al., 2001). In addition, Alu repeats are also involved in mRNA editing (Kreahling and Graveley, 2004), including mRNAs in the human brain (Blow et al., 2004). The functional significance of Alu repeats is consistent with the possibility of positive selection of these repeats in gene-rich regions of the human genome (International Human Genome Sequencing Consortium, 2001).

In addition, endogenous retroviruses (ERVs), the only long-terminal repeat (LTR) retrotransposons still active in the mammalian genome, cause several diseases (Druker and Whitelaw, 2004) and human endogenous retroviruses (HERVs) have been reported to influence the expression of some human genes (Britten, 1996), such as amylase genes (Samuelson et al., 1990; Ting et al., 1992). Interestingly, a type of HERV might be involved in the aetiology of schizophrenia, although the role of such inserts remains undefined (Deb-Rinker et al., 1999; Yolken et al., 2000; Karlsson et al., 2001; Deb-Rinker et al., 2002).

Some interspersed repeats in the upstream region of human genes are conserved in the upstream sequences of othologous genes in rodents and appear to be responsible for the expression of these genes, such as the essential *frataxin* gene implicated in ataxia (Greene et al., 2005). Although the promoter of the human *frataxin* gene contains the mammalian initiator region and another regulatory element, expression of the gene seems to be mainly driven by interspersed repeats in its promoter under the conditions of experimental tests. Reporter gene assays carried out with constructs lacking these repeats, one LINE and one SINE, showed a 40% reduction in expression compared to that observed with the full length construct (Greene et al., 2005). Surprisingly, the region upstream of the transcription start site is poorly conserved in rodents and similarities between the 1.2 Kb sequences upstream of the initiation codon in human and rodents are restricted to only two small regions located in the 220 bp 5'UTR. It is interesting that the first region, 46 bp long only, overlaps an ancient LINE element. Moreover, deletion of this region in reporter gene assays, in addition to the repeats mentioned above, resulted in a reduction of expression by another 40%, from 60% to 20%. Finally, electrophoretic mobility shift assays allowed the identification of two sequences of approximately 30 bp, each located in one of the two conserved regions, and none of these sequence contained known transcription factor binding sites (Greene et al., 2005).

Thus, it is unlikely that CSSs identified in the set of known genes in the two candidate regions for psychosis on chromosome 4 overlap unknown interspersed repeats, while the few CSSs overlapping known repeats might be functional, in particular because they appear to have been conserved over at least 300 million years.

Conserved non-coding sequences identified in the candidate regions have a high potential for function and could represent yet unidentified exons or regulatory regions. These sequences are therefore targets for experiments aiming to map transcripts and investigate their transcriptional regulation. These sequences are also targets for SNP detection and some were found to contain SNPs to prioritise in allelic association studies performed by the laboratory to identify variants associated with psychosis. In addition to genes encoding receptor for dopamine receptor D5 (DRD5), and other G protein-coupled receptors such as GPR78 and GPR125, the two minimal

regions contain other candidate genes, such as *cholecystokinin (CCK) A receptor* (*CCKAR*) and *phophatidylinositol 4-kinase type-II β* (*PI4K2B*) for which details of some CSSs were given above.

Cholecystokinin (CCK) is a neurotransmitter in intestinal and central neurones which inhibits food intake in a dose-dependent manner (Gibbs et al., 1973; Degen et al., 2001). CCK is secreted from intestinal cells in response to food intake and is important in the regulation of energy balance as a post-prandial satiety signal maybe mediated by CCKAR (Noble et al., 1999; Hillebrand et al., 2002). The *CCKAR* gene may contribute to susceptibility to diabetes (Marchal-Victorion et al., 2002), severe obesity (Miller et al., 1995; Funakoshi et al., 2000) and bulimia nervosa (Geracioti and Liddle, 1988; Klein and Walsh, 2004). CCK also plays an important role in the function of the central nervous system via its interaction with dopamine and other neurotransmitters such as serotonin and gamma-aminobutyric acid (Rehfeld, 2004). CCK is one of the most abundant neuropeptide in the human brain, in particular in cortical areas (Larsson and Rehfeld, 1979; Noble et al., 1999; Hillebrand et al., 2002). Furthermore, a transition in a transcription factor binding site in the promoter of the *CCK* gene may be associated with smoking (Comings et al., 2001), alcoholism (Harada et al., 1998), delirium tremens (Okubo et al., 2002), hallucinations in patients with Parkinson's disease (Wang et al., 2003), anxiety and panic disorder (Wang et al., 1998; Hattori et al., 2001; Rotzinger and Vaccarino, 2003), and schizophrenia (Wang et al., 2002). In addition, regulation of the expression of the *CCK* gene seems to involve glygogen synthase kinase 3 β (GSK3B) (Hansen et al., 2004). This is noteworthy since this enzyme mediates the action of mood stabilisers (Klein and Melton, 1996; Stambolic et al., 1996) and has been associated with affective disorders (Barrett et al., 2003; Benedetti et al., 2004a; Benedetti et al., 2004b; Benedetti et al., 2005). Despite a study reporting that a microsatellite in the upstream region of the *CCK* gene does not seem to be associated with mood disorders in Japanese (Hattori et al., 2002), these findings strongly suggest that the cholecystokinin system may be involved in psychiatric diseases.

Two CCK receptors have been identified (Kopin et al., 1992; Wank et al., 1992). CCKAR and CCKBR are receptor-coupled G proteins involved in phosphoinositids (PIs) break down and intracellular calcium mobilisation, and distinguished by their affinities to various forms of the CCK peptide (Kopin et al., 1992; Pisegna et al., 1992). Both receptors control CCK-induced release of dopamine in different brain areas (Crawley, 1991; Wang et al., 2003). CCKAR has also been found in dopaminergic neurones in areas of the brain implicated with schizophrenia (Hokfelt et al., 1980).

Association between polymorphisms in the upstream region, including rs6448456, in the 5'UTR of the human *CCKAR* gene at position –85 numbered from initiation codon ('SNP_68' in Figure 6.7a), have been associated with hallucinations during withdrawal in Japanese alcohol dependent men (Okubo et al., 2002), auditory hallucinations and other positive symptoms in SCZ (Wei and Hemmings, 1999; Tachikawa et al., 2000; Zhang et al., 2000) and SCZ (Tachikawa et al., 2001). Another variant, in intron 1, has been suggested to interact with a variant in the promoter of the *CCK* gene in visual hallucinations experienced by patients with Parkinson's disease (Wang et al., 2003). These variants were shown in Figure 6.7 and include rs1800857 (SNP_62), rs1800855 (SNP_63), rs6448456 (SNP_68), rs1799723 (SNP_69), rs1800908 (SNP_70). These associations need further examination as no major role for CCK related polymorphisms have been reported in some populations with panic disorder (Hamilton et al., 2001), SCZ (Bowen et al., 1998), and alcoholism (Ishiguro et al., 1999; Vanakoski et al., 2001). Although, none of the polymorphisms in the *CCKAR* gene have been reported to be functional, it is interesting that two of them, rs4349588 and rs6448456 lie in conserved sub-sequences identified here in species as distantly related to human, mouse and rat, as chicken, zebrafish and Fugu (Figure 6.7c and Figure 6.7d). Assuming that evolutionary constraints that retained these sequences reflect their function, it may be possible that the effect of the polymorphisms in these sequences were not detected by the analyses carried out, and thus would have small effect on gene expression. Alternatively, functional variants in close vicinity of these loci may exist but have not been identified yet.

PI4K2B is an enzyme involved in phospholipid metabolism (Minogue et al., 2001; Barylko et al., 2001; Balla et al., 2002). Lipid synthesis and degradation are involved in two cellular processes. First, phosphoinositids (PIs) play roles in cytoskeletal organisation and in the interactions between the membrane and protein machinery involved in vesicle budding and traffic (Rohrbough and Broadie, 2005). Second, PIs serve as intracellular second messengers mediating signal transduction by hormones and neurotransmitters into the cell (Rohrbough and Broadie, 2005). Cellular signalling and intracellular trafficking are activated independently and this requires spatial and temporal control of PI metabolism. This is mediated partly by PI kinases, which phosphorylates PIs in specific compartments of the cell, including the plasma membrane, endosomes, secretory granules, and the Golgi complex. The various functions of PIs are governed by various combinations of phosphorylation of the 3', 4' and 5' positions of the inositol head group (Rohrbough and Broadie, 2005). Phosphorylation of the 4' position of the inositol ring by PI 4-kinases, generating PI(4)P, seems to be the rate limiting step in the production of higher phosphorylated second messengers such as dyacylglycerol (DAG), and phosphorylated inositol derivatives such as inositol 1,4,5-triphosphate (IP3) or $PI(4,5)P_2$, which is able to anchor signalling complexes to membranes. The regulated production and distribution of PI(4)P is highly complex and dynamic, and variation in the distribution of PI(4)P varies with the cell type (Rohrbough and Broadie, 2005; Weixel et al., 2005).

Several genes involved in the synthesis or dephosphorylation of $PI(4,5)P_2$ lie in regions linked to BPAD, including 10p, 18q, 21q and 22q (Baron, 1997; Sklar, 2002). Some of these regions are also linked to SCZ, such as 10p, 18q and 22q (Lewis et al., 2003). A member of the PI 4-kinase family, *PI 4-kinase type III α* (*PI4KCA*), on 22q11 linked to BPAD and SCZ (Baron, 1997; Sklar, 2002; Lewis et al., 2003), contains a polymorphism at position –31 showing a trend for association to BPAD (Saito et al., 2003). Similarly, the promoter of *PI 3-kinase class 3* (*PIK3C3*), on 18q also linked to BPAD and SCZ (Baron, 1997; Sklar, 2002; Lewis et al., 2003), contains a SNP at –432 which affects binding of a transcription factor and has been associated with both BPAD and SCZ (Stopkova et al., 2004a; Duan et al., 2005). In addition, a microsatellite close to the exon 9-intron 9 splice donor site of a

member of the *PI(4)P 5-kinase family*, *PIP5K2A*, may also to be involved in BPAD and SCZ (Stopkova et al., 2003). Finally, mood stabilisers such as lithium and valproic acid target the glycogen synthase kinase 3-β (GSK3B) (Klein and Melton, 1996; Stambolic et al., 1996). Polymorphisms in this gene, including a SNP in its promoter, have also been associated with affective disorders (Barrett et al., 2003; Benedetti et al., 2004a; Benedetti et al., 2004b) and long-term response to lithium in BPAD (Benedetti et al., 2005).

The type II family of PI 4-kinases, comprising PI 4-kinase type II α (PI4K2A) and PI 4-kinase type II β (PI4K2B), have been isolated recently (Barylko et al., 2001; Minogue et al., 2001). PI4K2A and PI4K2B harbour little similarity to other protein or lipid kinases and represent a novel and distinct branch of the PI 4-kinase family (Barylko et al., 2001; Minogue et al., 2001). Rates of activity and kinetics of the two membrane-bound isoforms are identical, and the two enzymes have been suggested to have overlapping but not identical functions (Wei et al., 2002). PI4K2A appears to be an integral protein, while PI4K2B acts as a peripheral protein (Wei et al., 2002). Both enzymes have been reported to localise primarily to intracellular membranes in the endoplasmic reticulum, Golgi complex and in plasma membrane rafts (Wei et al., 2002; Waugh et al., 2003). PI4K2B is primarily cytosolic and associates with plasma membrane in higher levels than PI4K2A, while PI4K2A was also found in synaptic vesicles of neurones in the brain (Wei et al., 2002).

PI4K2B seems to be involved in the endocytotic pathway in the cytoplasm and in the production of second messengers when it is recruited to the plasma membrane (Heilmeyer et al., 2003). PI4K2B may contribute to the transmission of signals from some growth factors when it is recruited to the membrane and activated by AKT1 ('v-akt murine thymona viral oncogene homolog 1') (Wei et al., 2002). Interestingly, recent studies in human and mouse suggest that AKT1 may play a role in the aetiology of SCZ (Emamian et al., 2004; Ikeda et al., 2004). Furthermore, in response to neurotrophins, AKT1 inactivates GSK3B and may thereby promote axon outgrowth (Yoshimura et al., 2005). The *GSK3B* gene has also been associated with affective disorders (Barrett et al., 2003; Benedetti et al., 2004a; Benedetti et al., 2004b; Benedetti et al., 2005).

Type II PI 4-kinases colocalise in membrane lipid rafts with some growth factor receptors and may contribute to their endocytosis and processing (Minogue et al., 2001; Waugh et al., 2001). Lipid rafts are membrane microdomains rich in cholesterol and $PI(4,5)P_2$ that may serve as platforms for a range of cellular signalling complexes (Tsui-Pierchala et al., 2002). In neurones, lipid rafts are important for cell adhesion, axon guidance and synaptic transmission (Tsui-Pierchala et al., 2002). For example, the translocation of protein tyrosine kinases to lipid rafts is necessary for signalling and regulation of synaptic plasticity by neuregulin (Yang et al., 2004), which is implicated in SCZ (Stefansson et al., 2002). It is not known whether PI4K2B plays a role in neurones, but neuronal PI 4-kinase activity has been detected (Wiedemann et al., 1998). This seems to be accounted for mainly by PI4K2A, however, when it associates with synaptic vesicles (Guo et al., 2003). Nonetheless, PI4K2A activity does not account for the totality of PI 4-kinase activity (Guo et al., 2003), and it is possible that PI4K2B contributes to the generation of PI(4)P, hence of $PI(4,5)P_2$, in lipid rafts enabling them to function.

PI4K2B might also play a role in vesicles turn-over. Neurotransmitters are released in the synaptic space by exocytosis of the synaptic vesicles that contain them. The membrane of these vesicles then re-enter the cell by endocytosis and are recycled (Rohrbough and Broadie, 2005). Such transport of vesicles, including vesicles in synapses, relies on phosphoinositide synthesis and hydrolysis (Cremona and De Camilli, 2001; Rohrbough and Broadie, 2005). Generation of $PI(4,5)P_2$ to compensate its degradation during endocytosis is thought to occur on the plasma membrane via the action of PI(4)P 5-kinase type Iγ (PIP5K1C), using PI(4)P as a substrate (Wenk et al., 2001). Generation of PI(4)P itself is due to PI 4-kinase activity, which has been observed on synaptic vesicle (Wiedemann et al., 1998). The family of type II PI 4-kinases has been reported to account for an important proportion of PI 4-kinase activity, including activity on membranes, in various tissues (Barylko et al., 2001; Minogue et al., 2001). PI4K2A is present in neurones and is thought to account for the majority of the synaptic vesicle membrane-bound PI 4-kinase activity (Hinshaw, 2000; Guo et al., 2003). PI4K2A may contribute to synaptic vesicle cycle by generating PI(4)P from PI, thereby enabling creation of $PI(4,5)P_2$ in synaptic vesicles after its degradation by polyphosphoinositide

phosphatase synaptojanin 1 (SYNJ1), required for fusion of vesicles with the membrane (Guo et al., 2003). Although the major form of PI kinase activity in the brain seems to be PI 4-kinase activity and mediated by PI4K2A, it is possible that other PI 4-kinases play important roles in the brain upon activation (Zhao et al., 2001). Because PI4K2A and PI4K2B have very similar activities, PI4K2B may account for some of the PI 4-kinase activity observed in the brain. Interestingly, the *SYNJ1* gene lies on 21q22 linked to BPAD (Baron, 1997; Sklar, 2002) and two studies have reported trends of association of the gene with BPAD (Saito et al., 2001; Stopkova et al., 2004b). Evidence suggesting a potential role of type II PI 4-kinases in neuropsychiatric disease is also provided by a study reporting that activity of plasma membrane-bound type II PI 4-kinase activity in rat is inhibited directly and reversibly by pathophysiological concentrations of amyloid β proteins, and that type II PI 4-kinases may therefore be implicated in the pathogenesis of Alzheimer's disease (Wu et al., 2004).

The role of the recently identified *PI4K2B* gene in the PI metabolism is unclear at present. Because the type II PI 4-kinase family has recently been identified, the tissue and cellular distributions and the functions of both isoforms are not certain. Although it is not clear whether PI4K2B occurs in the brain, and in particular in neurones, it seems possible that this enzyme, like PI4K2A, is involved in the turn-over of synaptic vesicles. The function of PI4K2B may indeed be elusive in the cell types and conditions tested since PI4K2B appears to be mainly cytosolic and inactive, but activated upon recruitment to the plasma membrane. It is noteworthy that the *PI4K2B* gene lies on 4p15, linked to psychosis, and is involved in PI metabolism, a situation reminiscent of other genes encoding PI kinases that are located on chromosomal regions linked to BPAD and SCZ, and that have been associated with either or both of these illnesses, such as another PI 4-kinase, PI4KCA (Saito et al., 2003). *PI4K2B* thus appears to be another strong candidate gene in the 4p15 region. Regions identified in this gene that are conserved in distantly related species and were described above are likely to contribute to the function of PI4K2B and may be involved in susceptibility to BPAD and SCZ.

The examples shown here illustrate that genes in the candidate region, including strong candidate genes, contain non-coding sequences of 40 bp to 240 bp that are conserved in vertebrates as distantly related to mammals as bird and fish. Several conserved non-coding sequences contain SNPs, some of which were reported to be associated with neuropsychiatric diseases such as SCZ, and two of them also contain a putative binding site for the LIM Homeobox 3 (LXH3) transcription factor. These findings confirm the view that comparative analysis is a powerful tool for the identification of potentially functional non-coding regions, and can be applied to genes in a candidate region for psychosis to provide a limited number of sequences of particular interest.

CHAPTER 7

FINAL CONCLUSIONS

7 FINAL CONCLUSIONS

7.1 Summary

Affective disorders and schizophrenia are common and major psychiatric diseases that involve mood-related symptoms and disturbed thinking, respectively (Gelder et al., 2000). In addition to impairing the patient's life, these illnesses represent an important burden placed upon the patient's family, social and health services, and account for an substantial amount of the total NHS spending (Gelder et al., 2000). Although causes of these diseases are uncertain, they include genetic factors and seem to involve the serotonergic, dopaminergic and glutamatergic pathways (Baron, 1997; Harrison and Weinberger, 2005; Kato, 2001). Furthermore, treatment of these disorders through medication is often palliative and response to treatment varies widely between patients (Malhotra et al., 2004). Unravelling the aetiology of these diseases is therefore the focus of many academic research groups and pharmaceutical companies. Several candidate regions for bipolar affective disorder (BPAD) and schizophrenia (SCZ) have been identified by linkage studies (Owen and Cardno, 1999; Potash and DePaulo, 2000; Sklar, 2002). One of these regions was identified by the group on 4p15-16 (Blackwood et al., 1996). The group also set up an ACeDB database to manage the annotated genomic sequence of this region. The key activity of the group now consists in carrying out case-control association studies of single markers and haplotypes in large samples of control, unipolar (UP), BPAD and SCZ DNA. SNPs used in these studies are either retrieved from public resources, such as dbSNP (Smigielski et al., 2000), or detected in-house in tested samples. Due to the large number of markers to test and the large size of the case and control samples required to achieve statistical significance, DNA pooling is used in a fast, cost-effective initial screen of these markers. SNPs retained are then tested using classical association studies based on individual genotypes and haplotypes.

The first aim of this thesis was to develop a computational tool to assist ongoing laboratory work by allowing association data management and analysis. This was achieved by creating an extension of the ACeDB database and a CGI front-end to allow for the submission and analysis of allelic association data. Specifically, these two tools allow: i) storage of information relevant to genotyping assays (experimental conditions and results) and to DNA pool construction, ii) creation of meta data sets, and iii) test of allelic association with disease based on single DNA pools, single sets of individual samples in a population, or merged data sets, and storage of results of these tests. Since their development, these tools have been used successfully by experimental scientists with little knowledge of the intricacies of ACeDB, demonstrating their usefulness.

The goal of an association study is to detect differences in allele frequencies between affected and unaffected people that are expected to indicate differences in the function of genomic sequence variants at the loci studied. Markers tested by the group so far were located in coding sequences of genes in the two high-priority candidate regions B and D. Evidence is however accumulating that polymorphisms in non-coding sequences, in particular regulatory regions of genes, may play a role in the aetiology of complex disorders as important as those in coding sequences (Knight, 2005; Wray et al., 2003). Annotation of both coding and non-coding sequences of the candidate 4p region with functional elements, either known or novel, is therefore key to the identification of polymorphisms contributing to the diseases directly, or located close to causal variants. As coding sequences are estimated to represent only 2% of the mammalian genome (Rat Genome Sequencing Project Consortium, 2004), functional elements in non-coding sequences must be identified among a vast amount of sequence. Following the development of the database extension and CGI front-end to manage and analyse data generated by ongoing work, this thesis therefore set out to develop computational tools to identify functional elements in non-coding sequences.

Little is known about the types and quantities of functional elements in non-coding sequences, referred to until recently as 'junk' DNA. Regulatory regions of genes, which are located immediately upstream of genes, such as promoters, or within introns, or many kilobases away from the gene regulated, play a crucial role in

gene transcription regulation and have therefore been the focus of many studies of non-coding functional elements (Orphanides and Reinberg, 2002; Kadonaga, 2004; Svejstrup, 2004). These regions contain transcription factor binding sites (TFBSs) that are represented by small, between 6 and 20 bp, and often degenerate sequence motifs. Searching a sequence for these motifs typically generate a high number of false positives as the motifs are so small that they can occur by chance alone (Duret and Bucher, 1997). Restricting the search to sequences conserved between human and mouse can help address this problem (Tagle et al., 1988; Wasserman et al., 2000; Pennacchio and Rubin, 2001). The first method tested in this project aimed to identify putative regulatory regions by combining comparative genomics of upstream regions of genes in human and mouse with prediction of known transcription factor binding sites and detection of novel motifs in sequences conserved in the two species (Chapter 4). This approach was not sensitive enough and predicted too high a number of false positives to guide experimental work in the laboratory.

Most methods used to identify conserved human genomic sequences have compared human and mouse sequences and identified sub-sequences with a given minimum length and minimum percentage identity (Loots et al., 2000; Wasserman et al., 2000). Analysis of sequences from additional species reduced the amount of conserved neutral sequences, as divergence eliminates these sequences differently in each species and lineage (Duret and Bucher, 1997; Thomas et al., 2003). Most of these studies considered a limited number of genes within the same narrow chromosomal region (Gottgens et al., 2002; Thomas et al., 2003). For example, a study using sequences from 12 species and the scoring scheme based on the probability of the amount of conservation measured, given an estimated rate of neutral substitution, reported results for a 1.8 Mb region around the CFTR gene only, and relied on a broad estimate of the quantity of functional sequence in the entire human genome (Thomas et al., 2003). This study was difficult to compare with others because it included sequences from many species, contrasting with the majority of studies based two or three species only (Dubchak et al., 2000; Gottgens et al., 2002; Wasserman et al., 2000; Dubchak and Frazer, 2003). In contrast with the study of the CFTR region which used the evolutionary depth of the sequences analysed to distinguish between functional and neutral sequences, Kolbe and co-

workers (2004) used statistical properties of genomic sequences from human, mouse and rat of genes at various locations in the human genome and reported high accuracy with a 80% sensitivity, 85% specificity and 84% positive predictive value. This 3-way regulatory score has so far been developed for human, mouse and rat only.

The second approach applied in this project relied on the analysis of sequences of a human gene and of its orthologs in the five vertebrate species for which the genome sequence was publicly available at the time (Chapter 5). Four scoring schemes were benchmarked against a set of genes with known functional regulatory regions, identifying the most accurate: the score reflecting the probability of observing the amount of conservation measured, given the rate of synonymous substitution estimated for the gene studied (the 'binomial score'). Results showed that this approach was sensitive, since it detected a large proportion of known coding and regulatory genomic sequences (sensitivity of 78%), was highly specific (specificity of 97%) and showed a reasonable predictive power (positive predictive value of 60%). In comparison to the two studies mentioned above by Thomas *et al.* (2003) and Kolbe *et al.* (2004), the binomial score used in this thesis is as accurate and has the advantages of being applicable and tailored to each gene, or genomic region, unlike the study of the CFTR region (Thomas et al., 2003), and of including sequences from all species for which the genome sequence is available, such as chicken, zebrafish and pufferfish, unlike the 3-way regulatory score (Kolbe et al., 2004). The binomial score predictions may be regarded as conservative since positive control regions identified experimentally in the control sequences are likely to represent only a proportion of regions within these sequences that are actually involved in the regulation of gene expression in a given cell in different tissues, or areas of the brain, at different stages of development, or in response to various environmental stimuli (Werner, 2000). The limited number of conserved sequences detected using this scoring scheme could therefore realistically form the basis of experimental work.

The binomial score approach was consequently applied to genes in the 4p15-16 candidate region for psychosis. Conserved sequences were detected in 26 known and predicted genes supported by expression data, and their 10 Kb flanking regions,

by using orthologs from three sets of species: i) rodents only, ii) rodents and chicken, and iii) rodents, chicken and fish. Non-coding sequences conserved in the highest number of species are strongly supported putative functional sequences. Sequences with a high potential for function are targets for experimental verification of their function and for SNP detection. SNPs already identified in these regions are markers to test in allelic association studies using pooled DNA.

Examples of conserved sub-sequences (CSSs) identified in four genes in the 4p region using sequences from mammals, bird and fish, and a conservative set of parameters, were described in the Chapter 6. All exons in the four genes were at least partially overlapped by CSSs. In addition, numerous CSSs were identified in non-coding sequences of these genes. In particular, CSSs in introns and the upstream region of two candidate genes for psychosis were described in details. The first codes for a receptor for one of the most abundant neuropeptides in the human brain, cholecystokinin (CCK) A receptor (CCKAR) (Pisegna et al., 1992) and the second encodes an isoform of a type II PI 4-kinase potentially implicated in the synaptic vesicle cycle, phophatidylinositol 4-kinase type-II β (PI4K2B) (Minogue et al., 2001). In addition to being conserved in vertebrates as distantly related as mammals and fish, and hence interesting in their own right, some of these non-coding sequences contain polymorphisms previously reported to be associated with SCZ (Bowen et al., 1998), panic disorder (Hamilton et al., 2001), alcoholism (Ishiguro et al., 1999; Vanakoski et al., 2001), and hallucinations experienced by patients with Parkinson's disease (Wang et al., 2003).

7.2 Future directions

To continue this project, it would prove valuable to extend the analysis to the entire candidate region by applying this approach to blocks of conserved synteny rather than genes and their flanking sequences. The threshold used for the selection of top-scoring windows would be derived from the total length of known coding sequences in the human sequence spanned by the block tested. Also, the threshold used to detect sequences reported here was the exact proportion of coding region in the human sequence and is likely to be conservative since 5% of the mammalian genome is estimated to contain functional sequences whereas coding sequences only

account for half this amount (Rat Genome Sequencing Project Consortium, 2004). Although conservation of all coding sequences is not expected, a less stringent threshold could be used, depending on the length of the block, to allow for the selection of additional sequences with high potential for function. The analysis of blocks of conserved synteny from human and other species with the method presented here requires that sequences contain no rearrangement, nor error in the assembly because sequences were aligned using the global multiple alignment tool MLAGAN (Brudno et al., 2003). This issue could be addressed by using a local multiple alignment tool (Schwartz et al., 2003). The method presented here relies on a window scanning the alignment to which a score is given based on the amount of conservation observed for each species compared to human. It might therefore be valuable to try and score each column in the alignment, using a scoring scheme based on its information content, and derive sequences with high potential for function based on these columns' scores. This work could also serve as a basis for detection within the non-coding conserved regions of known regulatory motifs, or novel motifs, either individually or arranged in clusters, using software available to date (Frith et al. 2003). Identified motifs could then be searched for in the entire candidate region. It may also prove useful to search the human genome with conserved sequences identified in the candidate region 4p15-16, as they may suggest the presence of genes elsewhere in the genome that share common putative regulatory regions. It may also be useful to inspect genes located within other susceptibility regions for BPAD or schizophrenia, identify conserved non-coding sequences in these regions and compare these to that identified in the 4p15-16 region. This may allow the identification of sequence motifs shared by several conserved regions located in different places in the human genome.

Using the three sets of vertebrate species serves to identify sequences conserved in distant species such as fish, and also those not conserved in fish but in more closely related species such as chicken. However, some functional sequences important for biological functions specific to mammals, such as brain functions, may not be identified by the analysis including chicken and fish sequences. It follows that, although putative functional non-coding sequences detected using sequences from distant vertebrates such as fish or chicken ought to be studied, sequences identified in human and rodent only will likely enable the identification of non-coding functional

sequences and polymorphisms within them. On the other hand, a certain proportion of the sequences conserved in rodents may not be functional, but part of the eutherian core genome (Rat Genome Sequencing Project Consortium, 2004). It would therefore be valuable to complement this approach, based solely on comparative genomics, with sequences not only conserved but also predicted to be functional, based on sequence conservation in rodent combined with the analysis of the statistical properties of nucleotides in the aligned sequences to distinguish neutral sequences from functional ones (Kolbe et al., 2004). In addition, some of the conserved sequences might be filtered out by including sequences from additional mammalian species, such as dog, when the sequence of their genome becomes available.

Continuing efforts from international genome sequencing projects allow the regular improvement of the assemblies of the human, mouse and rat genomes and will result in the release of genomes from other vertebrate species. In particular, mammalian species are more appropriate than chicken and fish to use for this research project focusing on a human disease because analysis of several mammalian species is more informative (Thomas et al., 2003). The genome of the dog, for example, is due to be released in the very near future, followed by that of the cat, cow, pig and horse (http///www.genome.gov). The genome of the short-tail opossum will soon be released too, and is likely to provide valuable data, because of the intermediate position of marsupials on the evolutionary tree of species, between closely related mammals and distantly related vertebrate species such as chicken (Graves and Westerman, 2002). Also, identifying orthologs of genes from the 4p15-16 candidate region in these mammalian species is expected to be easier than in more distantly related species, and conservation of synteny may span longer regions.

The approach tested here could be extended to the other 30 genes located in the candidate region. Since these genes are not well characterised, identifying their orthologs in rodents, chicken and fish may prove difficult and require time-consuming manual inspection of the data. In any case, analysis of these 30 genes using the sequences available to date would complement the set of non-coding putative functional sequences generated here and provide a substantial amount of reliable reagents for experimental verification and SNP detection for the entire candidate region.

Comparative genomics has been shown to be an efficient tool to identify functional sequences (Thomas et al., 2003). It has however also been shown here that even though regulatory regions in human show a substantial amount of conservation in rodents, and species as distantly related as chicken and fish, not all human regulatory regions are conserved in these species. This is consistent with the fact that a third of TFBSs in human are not expected to be conserved in mouse (Dermitzakis and Clark, 2002). This issue could be addressed by comparing sequences from six primates species (Boffelli et al., 2003), if sequence data was available for these species. This is not yet the case however, as chimpanzee is the only primate species of which the genome sequence has been released (http://www.genome.gov). Until the moment when sequences from several primates are available publicly, regulatory motifs in the 4p15-16 region that are functional in human but not present in rodent might be identified by detecting shared, over-represented sequence motifs in upstream regions of genes involved in the same biochemical pathways as genes in the 4p15-16, which could be identified using GO annotation terms (http://www.geneontology.org) or of genes co-regulated with genes in the candidate genes, identified using expression data.

7.3 Conclusion

To conclude, the development of the extension of the ACeDB database and the CGI front-end enabled ongoing work by the group by allowing management and analysis of data generated by allelic association studies based on pooled DNA and individual genotypes. The development of a protocol using comparative genomics between vertebrate species generated reliable predictions to guide future laboratory work and forms the basis for the development of further computational tools to improve annotation of non-coding sequences in the candidate region. This thesis has therefore provided essential data for the identification of polymorphisms and haplotypes, particularly in non-coding sequences, which are associated with psychosis. Ultimately, this work will also contribute to the unravelling of the complex processes involved in normal brain function, and how their disruption leads to psychosis.

REFERENCES

8 REFERENCES

- Abdolmaleky,H.M., Smith,C.L., Faraone,S.V., Shafa,R., Stone,W., Glatt,S.J., and Tsuang,M.T. (2004). Methylomics in psychiatry: Modulation of gene-environment interactions may be through DNA methylation. Am. J. Med. Genet. B Neuropsychiatr. Genet. *127*, 51-59.
- Abecasis,G.R., Noguchi,E., Heinzmann,A., Traherne,J.A., Bhattacharyya,S., Leaves,N.I., Anderson,G.G., Zhang,Y., Lench,N.J., Carey,A., Cardon,L.R., Moffatt,M.F., and Cookson,W.O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. *68*, 191-197.
- Accili,D., Fishburn,C.S., Drago,J., Steiner,H., Lachowicz,J.E., Park,B.H., Gauda,E.B., Lee,E.J., Cool,M.H., Sibley,D.R., Gerfen,C.R., Westphal,H., and Fuchs,S. (1996). A targeted mutation of the D3 dopamine receptor gene is associated with hyperactivity in mice. Proc. Natl. Acad. Sci. U. S. A *93*, 1945-1949.
- Adams,F. (1856). The Extinct Works of Aretaeus, the Cappaocian (London: Sydenham Society Publications).
- Aebersold,R. and Goodlett,D.R. (2001). Mass spectrometry in proteomics. Chem. Rev. *101*, 269-295.
- Alexandersson,M., Cawley,S., and Pachter,L. (2003). SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res. *13*, 496-502.
- Alonso,C.R., Maxton-Kuechenmeister,J., and Akam,M. (2001). Evolution of Ftz protein function in insects. Curr. Biol. *11*, 1473-1478.
- Als,T.D., Dahl,H.A., Flint,T.J., Wang,A.G., Vang,M., Mors,O., Kruse,T.A., and Ewald,H. (2004). Possible evidence for a common risk locus for bipolar affective disorder and schizophrenia on chromosome 4p16 in patients from the Faroe Islands. Mol. Psychiatry *9*, 93-98.
- Altmuller,J., Palmer,L.J., Fischer,G., Scherb,H., and Wjst,M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. *69*, 936-950.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.
- Altshuler,D., Hirschhorn,J.N., Klannemark,M., Lindgren,C.M., Vohl,M.C., Nemesh,J., Lane,C.R., Schaffner,S.F., Bolk,S., Brewer,C., Tuomi,T., Gaudet,D., Hudson,T.J., Daly,M., Groop,L., and Lander,E.S. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet. *26*, 76-80.
- American Psychiatry Association (1994). Diagnostic and statistical manual of mental disorder, 4th edn. (Washington DC: American Association Press).
- Anand,S., Wang,W.C., Powell,D.R., Bolanowski,S.A., Zhang,J., Ledje,C., Pawashe,A.B., Amemiya,C.T., and Shashikant,C.S. (2003). Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. Proc. Natl. Acad. Sci. U. S. A *100*, 15666-15669.
- Anney,R.J., Rees,M.I., Bryan,E., Spurlock,G., Williams,N., Norton,N., Williams,H., Cardno,A., Zammit,S., Jones,S., Jones,G., Hoogendoorn,B., Smith,K., Hamshere,M.L., Coleman,S., Guy,C., O'Donovan,M.C., Owen,M.J., and Buckland,P.R. (2002). Characterisation, mutation detection, and association analysis of alternative promoters and 5' UTRs of the human dopamine D3 receptor gene in schizophrenia. Mol. Psychiatry *7*, 493-502.
- Antonarakis,S.E., Blouin,J.L., Lasseter,V.K., Gehrig,C., Radhakrishna,U., Nestadt,G., Housman,D.E., Kazazian,H.H., Kalman,K., Gutman,G., Fantino,E., Chandy,K.G., Gargus,J.J., and Pulver,A.E. (1999). Lack of linkage or association between schizophrenia and the polymorphic trinucleotide repeat within the KCNN3 gene on chromosome 1q21. Am. J. Med. Genet. *88*, 348-351.

- Antonarakis,S.E., Irkin,S.H., Cheng,T.C., Scott,A.F., Sexton,J.P., Trusko,S.P., Charache,S., and Kazazian,H.H., Jr. (1984). beta-Thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site. Proc. Natl. Acad. Sci. U. S. A *81*, 1154-1158.
- Anttila,S., Illi,A., Kampman,O., Mattila,K.M., Lehtimaki,T., and Leinonen,E. (2004). Interaction between NOTCH4 and catechol-O-methyltransferase genotypes in schizophrenia patients with poor response to typical neuroleptics. Pharmacogenetics *14*, 303-307.
- Anttila,S., Kampman,O., Illi,A., Roivas,M., Mattila,K.M., Lassila,V., Lehtimaki,T., and Leinonen,E. (2003). NOTCH4 gene promoter polymorphism is associated with the age of onset in schizophrenia. Psychiatr. Genet. *13*, 61-64.
- Aparicio,S., Morrison,A., Gould,A., Gilthorpe,J., Chaudhuri,C., Rigby,P., Krumlauf,R., and Brenner,S. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc. Natl. Acad. Sci. U. S. A *92*, 1684-1688.
- Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.m., Dehal,P., Christoffels,A., Rash,S., Hoon,S., Smit,A., Gelpke,M.D.S., Roach,J., Oh,T., Ho,I.Y., Wong,M., Detter,C., Verhoef,F., Predki,P., Tay,A., Lucas,S., Richardson,P., Smith,S.F., Clark,M.S., Edwards,Y.J.K., Doggett,N., Zharkikh,A., Tavtigian,S.V., Pruss,D., Barnstead,M., Evans,C., Baden,H., Powell,J., Glusman,G., Rowen,L., Hood,L., Tan,Y.H., Elgar,G., Hawkins,T., Venkatesh,B., Rokhsar,D., and Brenner,S. (2002). Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. Science *297*, 1301-1310.
- Aranyi,T., Kerjean,A., Toth,S., Mallet,J., Meloni,R., and Paldi,A. (2002). Paradoxical methylation of the tyrosine hydroxylase gene in mouse preimplantation embryos. Genomics *80*, 558-563.
- Arinami,T., Gao,M., Hamaguchi,H., and Toru,M. (1997). A functional polymorphism in the promoter region of the dopamine D2 receptor gene is associated with schizophrenia. Hum. Mol. Genet. *6*, 577-582.
- Arinami,T., Itokawa,M., Aoki,J., Shibuya,H., Ookubo,Y., Iwawaki,A., Ota,K., Shimizu,H., Hamaguchi,H., and Toru,M. (1996). Further association study on dopamine D2 receptor variant S311C in schizophrenia and affective disorders. Am. J. Med. Genet. *67*, 133-138.
- Arinami,T., Itokawa,M., Enguchi,H., Tagaya,H., Yano,S., Shimizu,H., Hamaguchi,H., and Toru,M. (1994). Association of dopamine D2 receptor molecular variant with schizophrenia. Lancet *343*, 703-704.
- Arinami,T., Ohtsuki,T., Takase,K., Shimizu,H., Yoshikawa,T., Horigome,H., Nakayama,J., and Toru,M. (2001). Screening for 22q11 deletions in a schizophrenia population. Schizophr. Res. *52*, 167-170.
- Arranz,M.J., Munro,J., Birkett,J., Bolonna,A., Mancama,D., Sodhi,M., Lesch,K.P., Meyer,J.F., Sham,P., Collier,D.A., Murray,R.M., and Kerwin,R.W. (2000). Pharmacogenetic prediction of clozapine response. Lancet *355*, 1615-1616.
- Arranz,M.J., Munro,J., Sham,P., Kirov,G., Murray,R.M., Collier,D.A., and Kerwin,R.W. (1998). Meta-analysis of studies on genetic variation in 5-HT2A receptors and clozapine response. Schizophr. Res. *32*, 93-99.
- Asherson,P., Mant,R., Williams,N., Cardno,A., Jones,L., Murphy,K., Collier,D.A., Nanko,S., Craddock,N., Morris,S., Muir,W., Blackwood,B., McGuffin,P., and Owen,M.J. (1998). A study of chromosome 4p markers and dopamine D5 receptor gene in schizophrenia and bipolar disorder. Mol. Psychiatry *3*, 310-320.
- Athanasiou,M.C., Malhotra,A.K., Xu,C., and Stephens,J.C. (2002). Discovery and utilization of haplotypes for pharmacogenetic studies of psychotropic drug response. Psychiatr. Genet. *12*, 89-96.
- Austin,C.P., Holder,D.J., Ma,L., Mixson,L.A., and Caskey,C.T. (1999). Mapping of hKCa3 to chromosome 1q21 and investigation of linkage of CAG repeat polymorphism to schizophrenia. Mol. Psychiatry *4*, 261-266.
- Badner,J.A. and Gershon,E.S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Mol. Psychiatry *7*, 405-411.
- Bailey,T.L. and Elkan,C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28-36.

- Baillarger (1854). De la folie à double forme. annales medico-psychologiques du système nerveux.
- Balla,A., Tuymetova,G., Barshishat,M., Geiszt,M., and Balla,T. (2002). Characterization of type II phosphatidylinositol 4-kinase isoforms reveals association of the enzymes with endosomal vesicular compartments. J. Biol. Chem. *277*, 20041-20050.
- Bannon,M.J., Michelhaugh,S.K., Wang,J., and Sacchetti,P. (2001). The human dopamine transporter gene: gene organization, transcriptional regulation, and potential involvement in neuropsychiatric disorders. Eur. Neuropsychopharmacol. *11*, 449-455.
- Barcellos,L.F., Klitz,W., Field,L.L., Tobias,R., Bowcock,A.M., Wilson,R., Nelson,M.P., Nagatomi,J., and Thomson,G. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. Am. J. Hum. Genet. *61*, 734-747.
- Barnes,N.M. and Sharp,T. (1999). A review of central 5-HT receptors and their function. Neuropharmacology *38*, 1083-1152.
- Baron,M. (1997). Genetic linkage and bipolar affective disorder: progress and pitfalls. Mol. Psychiatry *2*, 200-210.
- Barr,C.L., Wigg,K.G., Feng,Y., Zai,G., Malone,M., Roberts,W., Schachar,R., Tannock,R., and Kennedy,J.L. (2000). Attention-deficit hyperactivity disorder and the gene for the dopamine D5 receptor. Mol. Psychiatry *5*, 548-551.
- Barrett,T.B., Hauger,R.L., Kennedy,J.L., Sadovnick,A.D., Remick,R.A., Keck,P.E., McElroy,S.L., Alexander,M., Shaw,S.H., and Kelsoe,J.R. (2003). Evidence that a single nucleotide polymorphism in the promoter of the G protein receptor kinase 3 gene is associated with bipolar disorder. Mol. Psychiatry *8*, 546-557.
- Bartlett,S., Straub,J., Tonks,S., Wells,R.S., Bodmer,J.G., and Bodmer,W.F. (2001). Alkaline-mediated differential interaction (AMDI): a simple automatable single-nucleotide polymorphism assay. Proc. Natl. Acad. Sci. U. S. A *98*, 2694-2697.
- Barylko,B., Gerber,S.H., Binns,D.D., Grichine,N., Khvotchev,M., Sudhof,T.C., and Albanesi,J.P. (2001). A novel family of phosphatidylinositol 4-kinases conserved from yeast to humans. J. Biol. Chem. *276*, 7705-7708.
- Bassett,A.S., Hodgkinson,K., Chow,E.W., Correia,S., Scutt,L.E., and Weksberg,R. (1998). 22q11 deletion syndrome in adults with schizophrenia. Am. J. Med. Genet. *81*, 328-337.
- Bassett,A.S. and Husted,J. (1997). Anticipation or ascertainment bias in schizophrenia? Penrose's familial mental illness sample. Am. J. Hum. Genet. *60*, 630-637.
- Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B., and Lander,E.S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res. *10*, 950-958.
- Bayless,T.M., Picco,M.F., and LaBuda,M.C. (1998). Genetic anticipation in Crohn's disease. Am. J. Gastroenterol. *93*, 2322-2325.
- Baysal,B.E., Willett-Brozick,J.E., Badner,J.A., Corona,W., Ferrell,R.E., Nimgaonkar,V.L., and Detera-Wadleigh,S.D. (2002). A mannosyltransferase gene at 11q23 is disrupted by a translocation breakpoint that co-segregates with bipolar affective disorder in a small family. Neurogenetics *4*, 43-53.
- Bell,G.I., Horita,S., and Karam,J.H. (1984). A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes *33*, 176-183.
- Benedetti,F., Bernasconi,A., Lorenzi,C., Pontiggia,A., Serretti,A., Colombo,C., and Smeraldi,E. (2004a). A single nucleotide polymorphism in glycogen synthase kinase 3-beta promoter gene influences onset of illness in patients affected by bipolar disorder. Neurosci. Lett. *355*, 37-40.
- Benedetti,F., Serretti,A., Colombo,C., Lorenzi,C., Tubazio,V., and Smeraldi,E. (2004b). A glycogen synthase kinase 3-beta promoter gene single nucleotide polymorphism is associated with age at onset and response to total sleep deprivation in bipolar depression. Neurosci. Lett. *368*, 123-126.
- Benedetti,F., Serretti,A., Pontiggia,A., Bernasconi,A., Lorenzi,C., Colombo,C., and Smeraldi,E. (2005). Long-term response to lithium salts in bipolar illness is influenced by the glycogen synthase kinase 3-beta -50 T/C SNP. Neurosci. Lett. *376*, 51-55.

- Bentley,D.R. (2000). The Human Genome Project--an overview. Med. Res. Rev. *20*, 189-196.
- Bergh,S. and Cole,S.T. (1994). MycDB: an integrated mycobacterial database. Mol. Microbiol. *12*, 517-534.
- Berrettini,W.H. (2000). Are schizophrenic and bipolar disorders related? A review of family and molecular studies. Biol. Psychiatry *48*, 531-538.
- Berrettini,W. (2004). Bipolar disorder and schizophrenia: convergent molecular data. Neuromolecular. Med. *5*, 109-117.
- Bertelsen,A., Harvald,B., and Hauge,M. (1977). A Danish twin study of manic-depressive disorders. Br. J. Psychiatry *130*, 330-351.
- Bijl,R.V., de Graaf,R., Hiripi,E., Kessler,R.C., Kohn,R., Offord,D.R., Ustun,T.B., Vicente,B., Vollebergh,W.A., Walters,E.E., and Wittchen,H.U. (2003). The prevalence of treated and untreated mental disorders in five countries. Health Aff. *22*, 122-133.
- Blackwell,J.M. and Melville,S.E. (1999). Status of protozoan genome analysis: trypanosomatids. Parasitology *118 Suppl*, S11-S14.
- Blackwood,D.H., He,L., Morris,S.W., McLean,A., Whitton,C., Thomson,M., Walker,M.T., Woodburn,K., Sharp,C.M., Wright,A.F., Shibasaki,Y., St Clair,D.M., Porteous,D.J., and Muir,W.J. (1996). A locus for bipolar affective disorder on chromosome 4p. Nat. Genet. *12*, 427-430.
- Blackwood,D.H., Visscher,P.M., and Muir,W.J. (2001). Genetic studies of bipolar affective disorder in large families. Br. J. Psychiatry *178*, S134-S136.
- Blangero,J. (2004). Localization and identification of human quantitative trait loci: king harvest has surely come. Curr. Opin. Genet. Dev. *14*, 233-240.
- Bleuler, E. (1911). Dementia praecox oder Gruppe der Schizophrenien., In: G. Aschaffenburg (ed.), *Handbuch der Psychiatrie. Spezieller Teil. 4. Abteilung, 1.Hälfte.* (Leipzig: Franz Deuticke).
- Blits,K.C. (1999). Aristotle: form, function, and comparative anatomy. Anat. Rec. *257*, 58-63.
- Blouin,J.L., Dombroski,B.A., Nath,S.K., Lasseter,V.K., Wolyniec,P.S., Nestadt,G., Thornquist,M., Ullrich,G., McGrath,J., Kasch,L., Lamacz,M., Thomas,M.G., Gehrig,C., Radhakrishna,U., Snyder,S.E., Balk,K.G., Neufeld,K., Swartz,K.L., DeMarchi,N., Papadimitriou,G.N., Dikeos,D.G., Stefanis,C.N., Chakravarti,A., Childs,B., Housman,D.E., Kazazian,H.H., Antonarakis,S., and Pulver,A.E. (1998). Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. Nat. Genet. *20*, 70-73.
- Blow,M., Futreal,P.A., Wooster,R., and Stratton,M.R. (2004). A survey of RNA editing in human brain. Genome Res. *14*, 2379-2387.
- Bocchetta,A., Piccardi,M.P., Palmas,M.A., Chillotti,C., Oi,A., and Del Zompo,M. (1999). Family-based association study between bipolar disorder and DRD2, DRD4, DAT, and SERT in Sardinia. Am. J. Med. Genet. *88*, 522-526.
- Boffelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K.D., Ovcharenko,I., Pachter,L., and Rubin,E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*, 1391-1394.
- Boissinot,S., Entezam,A., Young,L., Munson,P.J., and Furano,A.V. (2004). The insertional history of an active family of L1 retrotransposons in humans. Genome Res. *14*, 1221-1231.
- Bonnet-Brilhault,F., Laurent,C., Campion,D., Thibaut,F., Lafargue,C., Charbonnier,F., Deleuze,J.F., Menard,J.F., Jay,M., Petit,M., Frebourg,T., and Mallet,J. (1999). No evidence for involvement of KCNN3 (hSKCa3) potassium channel gene in familial and isolated cases of schizophrenia. Eur. J. Hum. Genet. *7*, 247-250.
- Botstein,D., White,R.L., Skolnick,M., and Davis,R.W. (1980). Construction of a genetic linkage map in man using restriction fragment. Am. J. Hum. Genet. *32*, 314-331.
- Botstein,D. and Risch,N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33 Suppl*, 228-237.

- Botto,L.D., May,K., Fernhoff,P.M., Correa,A., Coleman,K., Rasmussen,S.A., Merritt,R.K., O'Leary,L.A., Wong,L.Y., Elixson,E.M., Mahle,W.T., and Campbell,R.M. (2003). A population-based study of the 22q11.2 deletion: phenotype, incidence, and contribution to major birth defects in the population. Pediatrics *112*, 101-107.
- Bowden,C., Theodorou,A.E., Cheetham,S.C., Lowther,S., Katona,C.L., Crompton,M.R., and Horton,R.W. (1997). Dopamine D1 and D2 receptor binding sites in brain samples from depressed suicides and controls. Brain Res. *752*, 227-233.
- Bowen,T., Guy,C.A., Craddock,N., Cardno,A.G., Williams,N.M., Spurlock,G., Murphy,K.C., Jones,L.A., Gray,M., Sanders,R.D., McCarthy,G., Chandy,K.G., Fantino,E., Kalman,K., Gutman,G.A., Gargus,J.J., Williams,J., McGuffin,P., Owen,M.J., and O'Donovan,M.C. (1998a). Further support for an association between a polymorphic CAG repeat in the hKCa3 gene and schizophrenia. Mol. Psychiatry *3*, 266-269.
- Bowen,T., Norton,N., Jacobsen,N.J., Guy,C., Daniels,J.K., Sanders,R.D., Cardno,A.G., Jones,L.A., Murphy,K.C., McGuffin,P., Craddock,N., O'Donovan,M.C., and Owen,M.J. (1998b). Linked polymorphisms upstream of exons 1 and 2 of the human cholecystokinin gene are not associated with schizophrenia or bipolar disorder. Mol. Psychiatry *3*, 67-71.
- Bowen,T., Williams,N., Norton,N., Spurlock,G., Wittekindt,O.H., Morris-Rosendahl,D.J., Williams,H., Brzustowicz,L., Hoogendoorn,B., Zammit,S., Jones,G., Sanders,R.D., Jones,L.A., McCarthy,G., Jones,S., Bassett,A., Cardno,A.G., Owen,M.J., and O'Donovan,M.C. (2001). Mutation screening of the KCNN3 gene reveals a rare frameshift mutation. Mol. Psychiatry *6*, 259-260.
- Boyer,P.A., Skolnick,P., and Fossom,L.H. (1998). Chronic administration of imipramine and citalopram alters the expression of NMDA receptor subunit mRNAs in mouse brain. A quantitative in situ hybridization study. J. Mol. Neurosci. *10*, 219-233.
- Brandon,N.J., Handford,E.J., Schurov,I., Rain,J.C., Pelling,M., Duran-Jimeniz,B., Camargo,L.M., Oliver,K.R., Beher,D., Shearman,M.S., and Whiting,P.J. (2004). Disrupted in Schizophrenia 1 and Nudel form a neurodevelopmentally regulated protein complex: implications for schizophrenia and other major neurological disorders. Mol. Cell Neurosci. *25*, 42-55.
- Bray,N. and Pachter,L. (2004). MAVID: constrained ancestral alignment of multiple sequences. Genome Res. *14*, 693-699.
- Bray,N.J., Buckland,P.R., Owen,M.J., and O'Donovan,M.C. (2003a). Cis-acting variation in the expression of a high proportion of genes in human brain. Hum. Genet. *113*, 149-153.
- Bray,N.J., Buckland,P.R., Williams,N.M., Williams,H.J., Norton,N., Owen,M.J., and O'Donovan,M.C. (2003b). A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. Am. J. Hum. Genet. *73*, 152-161.
- Brem,R.B., Yvert,G., Clinton,R., and Kruglyak,L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752-755.
- Brent,M.R. and Guigo,R. (2004). Recent advances in gene structure prediction. Curr. Opin. Struct. Biol. *14*, 264-272.
- Britten,R.J. (1994). Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc. Natl. Acad. Sci. U. S. A *91*, 5992-5996.
- Britten,R.J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. U. S. A *93*, 9374-9377.
- Britten,R.J. (1997). Mobile elements inserted in the distant past have taken on important functions. Gene *205*, 177-182.
- Brown,S., Fellers,J., Shippy,T., Denell,R., Stauber,M., and Schmidt-Ott,U. (2001). A strategy for mapping bicoid on the phylogenetic tree. Curr. Biol. *11*, R43-R44.
- Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A., and Batzoglou,S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. *13*, 721-731.
- Brunner,H.G., Nelen,M., Breakefield,X.O., Ropers,H.H., and van Oost,B.A. (1993). Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. Science *262*, 578-580.

- Brzustowicz,L.M., Hodgkinson,K.A., Chow,E.W., Honer,W.G., and Bassett,A.S. (2000). Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. Science *288*, 678-682.
- Buckland,P.R. (2004). Allele-specific gene expression differences in humans. Hum. Mol. Genet. *13 Spec No 2*, R255-R260.
- Bulyk,M.L. (2003). Computational prediction of transcription-factor binding site locations. Genome Biol. *5*, 201.
- Burda,P., te,H.S., Brachat,A., Wach,A., Dusterhoft,A., and Aebi,M. (1996). Stepwise assembly of the lipid-linked oligosaccharide in the endoplasmic reticulum of Saccharomyces cerevisiae: identification of the ALG9 gene encoding a putative mannosyl transferase. Proc. Natl. Acad. Sci. U. S. A *93*, 7160-7165.
- Burgert,E., Crocq,M.A., Bausch,E., Macher,J.P., and Morris-Rosendahl,D.J. (1998). No association between the tyrosine hydroxylase microsatellite marker HUMTH01 and schizophrenia or bipolar I disorder. Psychiatr. Genet. *8*, 45-48.
- Burren,O.S., Healy,B.C., Lam,A.C., Schuilenburg,H., Dolman,G.E., Everett,V.H., Laneri,D., Nutland,S., Rance,H.E., Payne,F., Smyth,D., Lowe,C., Barratt,B.J., Twells,R.C., Rainbow,D.B., Wicker,L.S., Todd,J.A., Walker,N.M., and Smink,L.J. (2004). Development of an integrated genome informatics, data management and workflow infrastructure: a toolbox for the study of complex disease genetics. Hum. Genomics *1*, 98-109.
- Burset,M. and Guigo,R. (1996). Evaluation of gene structure prediction programs. Genomics *34*, 353-367.
- Bussemaker,H.J., Li,H., and Siggia,E.D. (2001). Regulatory element detection using correlation with expression. Nat. Genet. *27*, 167-171.
- Buzdin,A., Ustyugova,S., Gogvadze,E., Lebedev,Y., Hunsmann,G., and Sverdlov,E. (2003). Genome-wide targeted search for human specific and polymorphic L1 integrations. Hum. Genet. *112*, 527-533.
- Byerley,W.F. (1989). Schizophrenia. Genetic linkage revisited. Nature *340*, 340-341.
- Caballero,I.M. and Hendrich,B. (2005). MeCP2 in neurons: closing in on the causes of Rett syndrome. Hum. Mol. Genet. *14 Spec No 1*, R19-R26.
- Caiafa,P. and Zampieri,M. (2005). DNA methylation and chromatin structure: the puzzling CpG islands. J. Cell Biochem. *94*, 257-265.
- Caldwell,C.B. and Gottesman,I.I. (1990). Schizophrenics kill themselves too: a review of risk factors for suicide. Schizophr. Bull. *16*, 571-589.
- Capper-Loup,C., Canales,J.J., Kadaba,N., and Graybiel,A.M. (2002). Concurrent activation of dopamine D1 and D2 receptors is required to evoke neural and behavioral phenotypes of cocaine sensitization. J. Neurosci. *22*, 6218-6227.
- Cardno,A.G., Bowen,T., Guy,C.A., Jones,L.A., McCarthy,G., Williams,N.M., Murphy,K.C., Spurlock,G., Gray,M., Sanders,R.D., Craddock,N., McGuffin,P., Owen,M.J., and O'Donovan,M.C. (1999). CAG repeat length in the hKCa3 gene and symptom dimensions in schizophrenia. Biol. Psychiatry *45*, 1592-1596.
- Cardon,L.R. and Bell,J.I. (2001). Association study designs for complex diseases. Nat. Rev. Genet. *2*, 91-99.
- Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N., Nemesh,J., Ziaugra,L., Friedland,L., Rolfe,A., Warrington,J., Lipshutz,R., Daley,G.Q., and Lander,E.S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. *22*, 231-238.
- Carlson,C., Papolos,D., Pandita,R.K., Faedda,G.L., Veit,S., Goldberg,R., Shprintzen,R., Kucherlapati,R., and Morrow,B. (1997). Molecular analysis of velo-cardio-facial syndrome patients with psychiatric disorders. Am. J. Hum. Genet. *60*, 851-859.
- Caron,H., van Schaik,B., van der,M.M., Baas,F., Riggins,G., van Sluis,P., Hermus,M.C., van Asperen,R., Boon,K., Voute,P.A., Heisterkamp,S., van Kampen,A., and Versteeg,R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science *291*, 1289-1292.

- Carroll,M.L., Roy-Engel,A.M., Nguyen,S.V., Salem,A.H., Vogel,E., Vincent,B., Myers,J., Ahmad,Z., Nguyen,L., Sammarco,M., Watkins,W.S., Henke,J., Makalowski,W., Jorde,L.B., Deininger,P.L., and Batzer,M.A. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J. Mol. Biol. *311*, 17-40.
- Cartmell,J. and Schoepp,D.D. (2000). Regulation of neurotransmitter release by metabotropic glutamate receptors. J. Neurochem. *75*, 889-907.
- Cases,O., Seif,I., Grimsby,J., Gaspar,P., Chen,K., Pournin,S., Muller,U., Aguet,M., Babinet,C., and Shih,J.C. (1995). Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. Science *268*, 1763-1766.
- Catalano,M. (2001). Functionally gene-linked polymorphic regions and genetically controlled neurotransmitters metabolism. Eur. Neuropsychopharmacol. *11*, 431-439.
- Caubit,X., Thangarajah,R., Theil,T., Wirth,J., Nothwang,H.G., Ruther,U., and Krauss,S. (1999). Mouse Dac, a novel nuclear factor with homology to Drosophila dachshund shows a dynamic expression in the neural crest, the eye, the neocortex, and the limb bud. Dev. Dyn. *214*, 66-80.
- Chakravarti,A. (1999). Population genetics--making sense out of sequence. Nat. Genet. *21*, 56-60.
- Chakravarti,A., Buetow,K.H., Antonarakis,S.E., Waber,P.G., Boehm,C.D., and Kazazian,H.H. (1984). Nonuniform recombination within the human beta-globin gene cluster. Am. J. Hum. Genet. *36*, 1239-1258.
- Chandy,K.G., Fantino,E., Wittekindt,O., Kalman,K., Tong,L.L., Ho,T.H., Gutman,G.A., Crocq,M.A., Ganguli,R., Nimgaonkar,V., Morris-Rosendahl,D.J., and Gargus,J.J. (1998). Isolation of a novel potassium channel gene hSKCa3 containing a polymorphic CAG repeat: a candidate for schizophrenia and bipolar disorder? Mol. Psychiatry *3*, 32-37.
- Chang,F.M., Kidd,J.R., Livak,K.J., Pakstis,A.J., and Kidd,K.K. (1996). The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus. Hum. Genet. *98*, 91-101.
- Chapman,M.A., Donaldson,I.J., Gilbert,J., Grafham,D., Rogers,J., Green,A.R., and Gottgens,B. (2004). Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. Genome Res. *14*, 313-318.
- Chen,J., Lipska,B.K., Halim,N., Ma,Q.D., Matsumoto,M., Melhem,S., Kolachana,B.S., Hyde,T.M., Herman,M.M., Apud,J., Egan,M.F., Kleinman,J.E., and Weinberger,D.R. (2004a). Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. Am. J. Hum. Genet. *75*, 807-821.
- Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.K., Chen,W.J., Cunningham,F., Davis,P., Kenny,E., Kishore,R., Lawson,D., Lee,R., Muller,H.M., Nakamura,C., Pai,S., Ozersky,P., Petcherski,A., Rogers,A., Sabo,A., Schwarz,E.M., Van Auken,K., Wang,Q., Durbin,R., Spieth,J., Sternberg,P.W., and Stein,L.D. (2005). WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res. *33*, D383-D389.
- Chen,W.Y., Shi,Y.Y., Zheng,Y.L., Zhao,X.Z., Zhang,G.J., Chen,S.Q., Yang,P.D., and He,L. (2004b). Case-control study and transmission disequilibrium test provide consistent evidence for association between schizophrenia and genetic variation in the 22q11 gene ZDHHC8. Hum. Mol. Genet. *13*, 2991-2995.
- Chen,X., Wang,X., O'Neill,A.F., Walsh,D., and Kendler,K.S. (2004c). Variants in the catechol-o-methyltransferase (COMT) gene are associated with schizophrenia in Irish high-density families. Mol. Psychiatry *9*, 962-967.
- Chennathukuzhi,V., Stein,J.M., Abel,T., Donlon,S., Yang,S., Miller,J.P., Allman,D.M., Simmons,R.A., and Hecht,N.B. (2003). Mice deficient for testis-brain RNA-binding protein exhibit a coordinate loss of TRAX, reduced fertility, altered gene expression in the brain, and behavioral changes. Mol. Cell Biol. *23*, 6419-6434.

- Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S., and Botstein,D. (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res. *26*, 73-79.
- Cheung,V.G., Conlin,L.K., Weber,T.M., Arcaro,M., Jen,K.Y., Morley,M., and Spielman,R.S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. Nat. Genet. *33*, 422-425.
- Cho,K.S., Elizondo,L.I., and Boerkoel,C.F. (2004). Advances in chromatin remodeling and human disease. Curr. Opin. Genet. Dev. *14*, 308-315.
- Chow,E.W., Mikulis,D.J., Zipursky,R.B., Scutt,L.E., Weksberg,R., and Bassett,A.S. (1999). Qualitative MRI findings in adults with 22q11 deletion syndrome and schizophrenia. Biol. Psychiatry *46*, 1436-1442.
- Chow,E.W., Zipursky,R.B., Mikulis,D.J., and Bassett,A.S. (2002). Structural brain abnormalities in patients with schizophrenia and 22q11 deletion syndrome. Biol. Psychiatry *51*, 208-215.
- Chowdari,K.V., Mirnics,K., Semwal,P., Wood,J., Lawrence,E., Bhatia,T., Deshpande,S.N., B K T, Ferrell,R.E., Middleton,F.A., Devlin,B., Levitt,P., Lewis,D.A., and Nimgaonkar,V.L. (2002). Association and linkage analyses of RGS4 polymorphisms in schizophrenia. Hum. Mol. Genet. *11*, 1373-1380.
- Chowdari,K.V., Wood,J., Ganguli,R., Gottesman,I.I., and Nimgaonkar,V.L. (2000). Lack of association between schizophrenia and a CAG repeat polymorphism of the hSKCa3 gene in a north eastern US sample. Mol. Psychiatry *5*, 237-238.
- Chumakov,I., Blumenfeld,M., Guerassimenko,O., Cavarec,L., Palicio,M., Abderrahim,H., Bougueleret,L., Barry,C., Tanaka,H., La Rosa,P., Puech,A., Tahri,N., Cohen-Akenine,A., Delabrosse,S., Lissarrague,S., Picard,F.P., Maurice,K., Essioux,L., Millasseau,P., Grel,P., Debailleul,V., Simon,A.M., Caterina,D., Dufaure,I., Malekzadeh,K., Belova,M., Luan,J.J., Bouillot,M., Sambucy,J.L., Primas,G., Saumier,M., Boubkiri,N., Martin-Saumier,S., Nasroune,M., Peixoto,H., Delaye,A., Pinchot,V., Bastucci,M., Guillou,S., Chevillon,M., Sainz-Fuertes,R., Meguenni,S., Aurich-Costa,J., Cherif,D., Gimalac,A., Van Duijn,C., Gauvreau,D., Ouellette,G., Fortier,I., Raelson,J., Sherbatich,T., Riazanskaia,N., Rogaev,E., Raeymaekers,P., Aerssens,J., Konings,F., Luyten,W., Macciardi,F., Sham,P.C., Straub,R.E., Weinberger,D.R., Cohen,N., and Cohen,D. (2002). Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. Proc. Natl. Acad. Sci. U. S. A *99*, 13675-13680.
- Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., and Birney,E. (2003). Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res. *31*, 38-42.
- Clayton,D. and McKeigue,P.M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet *358*, 1356-1360.
- Cohen,B.A., Mitra,R.D., Hughes,J.D., and Church,G.M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. *26*, 183-186.
- Cohen,S.M. and Nadler,J.V. (1997). Proline-induced potentiation of glutamate transmission. Brain Res. *761*, 271-282.
- Collier,D.A. and Li,T. (2003). The genetics of schizophrenia: glutamate not dopamine? Eur. J. Pharmacol. *480*, 177-184.
- Collier,D.A., Stober,G., Li,T., Heils,A., Catalano,M., Di Bella,D., Arranz,M.J., Murray,R.M., Vallada,H.P., Bengel,D., Muller,C.R., Roberts,G.W., Smeraldi,E., Kirov,G., Sham,P., and Lesch,K.P. (1996). A novel functional polymorphism within the promoter of the serotonin transporter gene: possible role in susceptibility to affective disorders. Mol. Psychiatry *1*, 453-460.

- Collins,F.S. (1995). Positional cloning moves from perditional to traditional. Nat. Genet. *9*, 347-350.
- Collins,F.S., Green,E.D., Guttmacher,A.E., and Guyer,M.S. (2003). A vision for the future of genomics research. Nature *422*, 835-847.
- Collins,F.S., Guyer,M.S., and Charkravarti,A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science *278*, 1580-1581.
- Comings,D.E., Wu,S., Gonzalez,N., Iacono,W.G., McGue,M., Peters,W.W., and MacMurray,J.P. (2001). Cholecystokinin (CCK) gene as a possible risk factor for smoking: a replication in two independent samples. Mol. Genet. Metab. *73*, 349-353.
- Conroy,J., Meally,E., Kearney,G., Fitzgerald,M., Gill,M., and Gallagher,L. (2004). Serotonin transporter gene and autism: a haplotype analysis in an Irish autistic population. Mol. Psychiatry *9*, 587-593.
- Corfas,G., Roy,K., and Buxbaum,J.D. (2004). Neuregulin 1-erbB signaling and the molecular/cellular basis of schizophrenia. Nat. Neurosci. *7*, 575-580.
- Corvin,A.P., Morris,D.W., McGhee,K., Schwaiger,S., Scully,P., Quinn,J., Meagher,D., Clair,D.S., Waddington,J.L., and Gill,M. (2004). Confirmation and refinement of an 'at-risk' haplotype for schizophrenia suggests the EST cluster, Hs.97362, as a potential susceptibility gene at the Neuregulin-1 locus. Mol. Psychiatry *9*, 208-213.
- Cousin,X., Hotelier,T., Giles,K., Toutant,J.P., and Chatonnet,A. (1998). aCHEdb: the database system for ESTHER, the alpha/beta fold family of proteins and the Cholinesterase gene server. Nucleic Acids Res. *26*, 226-228.
- Cowles,C.R., Hirschhorn,J.N., Altshuler,D., and Lander,E.S. (2002). Detection of regulatory variation in mouse genes. Nat. Genet. *32*, 432-437.
- Craddock,N., Daniels,J., Roberts,E., Rees,M., McGuffin,P., and Owen,M.J. (1995). No evidence for allelic association between bipolar disorder and monoamine oxidase A gene polymorphisms. Am. J. Med. Genet. *60*, 322-324.
- Craddock,N., McKeon,P., Moorhead,S., Guy,C., Harrison,D., Mynett-Johnson,L., Claffey,E., Feldman,E., McGuffin,P., Owen,M.J., and O'Donovan,M.C. (1997). Expanded CAG/CTG repeats in bipolar disorder: no correlation with phenotypic measures of illness severity. Biol. Psychiatry *42*, 876-881.
- Cramer,P. (2004). Structure and function of RNA polymerase II. Adv. Protein Chem. *67*, 1-42.
- Cravchik,A. and Gejman,P.V. (1999). Functional analysis of the human D5 dopamine receptor missense and nonsense variants: differences in dopamine binding affinities. Pharmacogenetics *9*, 199-206.
- Crawley,J.N. (1991). Cholecystokinin-dopamine interactions. Trends Pharmacol. Sci. *12*, 232-236.
- Cremona,O. and De Camilli,P. (2001). Phosphoinositides in membrane traffic at the synapse. J. Cell Sci. *114*, 1041-1052.
- Crocq,M.A., Mant,R., Asherson,P., Williams,J., Hode,Y., Mayerova,A., Collier,D., Lannfelt,L., Sokoloff,P., and Schwartz,J.C. (1992). Association between schizophrenia and homozygosity at the dopamine D3 receptor gene. J. Med. Genet. *29*, 858-860.
- Crollius,H.R., Jaillon,O., Dasilva,C., Ozouf-Costaz,C., Fizames,C., Fischer,C., Bouneau,L., Billault,A., Quetier,F., Saurin,W., Bernot,A., and Weissenbach,J. (2000). Characterization and Repeat Analysis of the Compact Genome of the Freshwater Pufferfish Tetraodon nigroviridis. Genome Res. *10*, 939-949.
- Curran,S., Purcell,S., Craig,I., Asherson,P., and Sham,P. (2005). The serotonin transporter gene as a QTL for ADHD. Am. J. Med. Genet. B Neuropsychiatr. Genet. *134*, 42-47.
- Cusin,C., Serretti,A., Lattuada,E., Lilli,R., Lorenzi,C., and Smeraldi,E. (2002a). Association study of MAO-A, COMT, 5-HT2A, DRD2, and DRD4 polymorphisms with illness time course in mood disorders. Am. J. Med. Genet. *114*, 380-390.

- Cusin,C., Serretti,A., Zanardi,R., Lattuada,E., Rossini,D., Lilli,R., Lorenzi,C., and Smeraldi,E. (2002b). Influence of monoamine oxidase A and serotonin receptor 2A polymorphisms in SSRI antidepressant activity. Int. J. Neuropsychopharmacol. *5*, 27-35.
- D'Souza,U.M., Russ,C., Tahir,E., Mill,J., McGuffin,P., Asherson,P.J., and Craig,I.W. (2004). Functional effects of a tandem duplication polymorphism in the 5'flanking region of the DRD4 gene. Biol. Psychiatry *56*, 691-697.
- Daggett,V. and Fersht,A. (2003). The present view of the mechanism of protein folding. Nat. Rev. Mol. Cell Biol. *4*, 497-502.
- Daly,G., Hawi,Z., Fitzgerald,M., and Gill,M. (1999). Mapping susceptibility loci in attention deficit hyperactivity disorder: preferential transmission of parental alleles at DAT1, DBH and DRD5 to affected children. Mol. Psychiatry *4*, 192-196.
- Davies,G., Welham,J., Chant,D., Torrey,E.F., and McGrath,J. (2003). A systematic review and meta-analysis of Northern Hemisphere season of birth studies in schizophrenia. Schizophr. Bull. *29*, 587-593.
- Davies,K. (2001). After the genome: DNA and human disease. Cell *104*, 465-467.
- Davis,P.K. and Brackmann,R.K. (2003). Chromatin remodeling and cancer. Cancer Biol. Ther. *2*, 22-29.
- Davis,R.J., Shen,W., Heanue,T.A., and Mardon,G. (1999). Mouse Dach, a homologue of Drosophila dachshund, is expressed in the developing retina, brain and limbs. Dev. Genes Evol. *209*, 526-536.
- Davuluri,R.V., Grosse,I., and Zhang,M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. *29*, 412-417.
- Dawson,E., Abecasis,G.R., Bumpstead,S., Chen,Y., Hunt,S., Beare,D.M., Pabial,J., Dibling,T., Tinsley,E., Kirby,S., Carter,D., Papaspyridonos,M., Livingstone,S., Ganske,R., Lohmussaar,E., Zernant,J., Tonisson,N., Remm,M., Magi,R., Puurand,T., Vilo,J., Kurg,A., Rice,K., Deloukas,P., Mott,R., Metspalu,A., Bentley,D.R., Cardon,L.R., and Dunham,I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature *418*, 544-548.
- Dawson,L.A. and Routledge,C. (1995). Differential effects of potassium channel blockers on extracellular concentrations of dopamine and 5-HT in the striatum of conscious rats. Br. J. Pharmacol. *116*, 3260-3264.
- Day,R., Nielsen,J.A., Korten,A., Ernberg,G., Dube,K.C., Gebhart,J., Jablensky,A., Leon,C., Marsella,A., and Olatawura,M. (1987). Stressful life events preceding the acute onset of schizophrenia: a cross-national study from the World Health Organization. Cult. Med. Psychiatry *11*, 123-205.
- De Blasi,A., Conn,P.J., Pin,J., and Nicoletti,F. (2001). Molecular determinants of metabotropic glutamate receptor signaling. Trends Pharmacol. Sci. *22*, 114-120.
- De Bruyn,A., Mendelbaum,K., Sandkuijl,L.A., Delvenne,V., Hirsch,D., Staner,L., Mendlewicz,J., and Van Broeckhoven,C. (1994). Nonlinkage of bipolar illness to tyrosine hydroxylase, tyrosinase, and D2 and D4 dopamine receptor genes on chromosome 11. Am. J. Psychiatry *151*, 102-106.
- De Vries,L., Zheng,B., Fischer,T., Elenko,E., and Farquhar,M.G. (2000). The regulator of G protein signaling family. Annu. Rev. Pharmacol. Toxicol. *40*, 235-271.
- Deb-Rinker,P., Klempan,T.A., O'Reilly,R.L., Torrey,E.F., and Singh,S.M. (1999). Molecular characterization of a MSRV-like sequence identified by RDA from monozygotic twin pairs discordant for schizophrenia. Genomics *61*, 133-144.
- Deb-Rinker,P., O'Reilly,R.L., Torrey,E.F., and Singh,S.M. (2002). Molecular characterization of a 2.7-kb, 12q13-specific, retroviral-related sequence isolated by RDA from monozygotic twin pairs discordant for schizophrenia. Genome *45*, 381-390.
- Deckert,J., Catalano,M., Syagailo,Y.V., Bosi,M., Okladnova,O., Di Bella,D., Nothen,M.M., Maffei,P., Franke,P., Fritze,J., Maier,W., Propping,P., Beckmann,H., Bellodi,L., and Lesch,K.P. (1999). Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder. Hum. Mol. Genet. *8*, 621-624.

- Deeb,S.S., Fajas,L., Nemoto,M., Pihlajamaki,J., Mykkanen,L., Kuusisto,J., Laakso,M., Fujimoto,W., and Auwerx,J. (1998). A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. Nat. Genet. *20*, 284-287.
- Degen,L., Matzinger,D., Drewe,J., and Beglinger,C. (2001). The effect of cholecystokinin in controlling appetite and food intake in humans. Peptides *22*, 1265-1269.
- Degrave,W., de Miranda,A.B., Amorim,A., Brandao,A., Aslett,M., and Vandeyar,M. (1997). TcruziDB, an integrated database, and the WWW information server for the Trypanosoma cruzi genome project. Mem. Inst. Oswaldo Cruz *92*, 805-809.
- Dehal,P., Predki,P., Olsen,A.S., Kobayashi,A., Folta,P., Lucas,S., Land,M., Terry,A., Ecale Zhou,C.L., Rash,S., Zhang,Q., Gordon,L., Kim,J., Elkin,C., Pollard,M.J., Richardson,P., Rokhsar,D., Uberbacher,E., Hawkins,T., Branscomb,E., and Stubbs,L. (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. Science *293*, 104-111.
- Deininger,P.L. and Batzer,M.A. (1999). Alu repeats and human disease. Mol. Genet. Metab *67*, 183-193.
- Deininger,P.L. and Batzer,M.A. (2002). Mammalian retroelements. Genome Res. *12*, 1455-1465.
- Delcher,A.L., Phillippy,A., Carlton,J., and Salzberg,S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. *30*, 2478-2483.
- DeLisi,L.E., Shaw,S.H., Crow,T.J., Shields,G., Smith,A.B., Larach,V.W., Wellman,N., Loftus,J., Nanthakumar,B., Razi,K., Stewart,J., Comazzi,M., Vita,A., Heffner,T., and Sherrington,R. (2002). A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. Am. J. Psychiatry *159*, 803-812.
- Demyttenaere,K., Bruffaerts,R., Posada-Villa,J., Gasquet,I., Kovess,V., Lepine,J.P., Angermeyer,M.C., Bernert,S., de Girolamo,G., Morosini,P., Polidori,G., Kikkawa,T., Kawakami,N., Ono,Y., Takeshima,T., Uda,H., Karam,E.G., Fayyad,J.A., Karam,A.N., Mneimneh,Z.N., Medina-Mora,M.E., Borges,G., Lara,C., de Graaf,R., Ormel,J., Gureje,O., Shen,Y., Huang,Y., Zhang,M., Alonso,J., Haro,J.M., Vilagut,G., Bromet,E.J., Gluzman,S., Webb,C., Kessler,R.C., Merikangas,K.R., Anthony,J.C., Von Korff,M.R., Wang,P.S., Brugha,T.S., Aguilar-Gaxiola,S., Lee,S., Heeringa,S., Pennell,B.E., Zaslavsky,A.M., Ustun,T.B., and Chatterji,S. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA *291*, 2581-2590.
- Dermitzakis,E.T. and Clark,A.G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. *19*, 1114-1121.
- Dermitzakis,E.T., Kirkness,E., Schwarz,S., Birney,E., Reymond,A., and Antonarakis,S.E. (2004). Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. Genome Res. *14*, 852-859.
- DeSilva,U., Elnitski,L., Idol,J.R., Doyle,J.L., Gan,W., Thomas,J.W., Schwartz,S., Dietrich,N.L., Beckstrom-Sternberg,S.M., McDowell,J.C., Blakesley,R.W., Bouffard,G.G., Thomas,P.J., Touchman,J.W., Miller,W., and Green,E.D. (2002). Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. Genome Res. *12*, 3-15.
- Detera-Wadleigh,S.D., Badner,J.A., Berrettini,W.H., Yoshikawa,T., Goldin,L.R., Turner,G., Rollins,D.Y., Moses,T., Sanders,A.R., Karkera,J.D., Esterling,L.E., Zeng,J., Ferraro,T.N., Guroff,J.J., Kazuba,D., Maxwell,M.E., Nurnberger,J.I., Jr., and Gershon,E.S. (1999). A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. Proc. Natl. Acad. Sci. U. S. A *96*, 5604-5609.
- Dever,T.E. (2002). Gene-specific regulation by general translation factors. Cell *108*, 545-556.

- Diehl,D.J. and Gershon,S. (1992). The role of dopamine in mood disorders. Compr. Psychiatry *33*, 115-120.
- Dillon,N. (2004). Heterochromatin structure and function. Biol. Cell *96*, 631-637.
- Dimitrova,A., Milanova,V., Krastev,S., Nikolov,I., Toncheva,D., Owen,M.J., and Kirov,G. (2005). Association study of myo-inositol monophosphatase 2 (IMPA2) polymorphisms with bipolar affective disorder and response to lithium treatment. Pharmacogenomics. J. *5*, 35-41.
- Dorman,J.S., LaPorte,R.E., Stone,R.A., and Trucco,M. (1990). Worldwide differences in the incidence of type I diabetes are associated with amino acid variation at position 57 of the HLA-DQ beta chain. Proc. Natl. Acad. Sci. U. S. A *87*, 7370-7374.
- Down,T.A. and Hubbard,T.J. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. *12*, 458-461.
- Drevets,W.C., Gautier,C., Price,J.C., Kupfer,D.J., Kinahan,P.E., Grace,A.A., Price,J.L., and Mathis,C.A. (2001). Amphetamine-induced dopamine release in human ventral striatum correlates with euphoria. Biol. Psychiatry *49*, 81-96.
- Drevets,W.C., Price,J.L., Simpson,J.R., Jr., Todd,R.D., Reich,T., Vannier,M., and Raichle,M.E. (1997). Subgenual prefrontal cortex abnormalities in mood disorders. Nature *386*, 824-827.
- Dror,V., Shamir,E., Ghanshani,S., Kimhi,R., Swartz,M., Barak,Y., Weizman,R., Avivi,L., Litmanovitch,T., Fantino,E., Kalman,K., Jones,E.G., Chandy,K.G., Gargus,J.J., Gutman,G.A., and Navon,R. (1999). hKCa3/KCNN3 potassium channel gene: association of longer CAG repeats with schizophrenia in Israeli Ashkenazi Jews, expression in human tissues and localization to chromosome 1q21. Mol. Psychiatry *4*, 254-260.
- Druker,R., Bruxner,T.J., Lehrbach,N.J., and Whitelaw,E. (2004). Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. Nucleic Acids Res. *32*, 5800-5808.
- Druker,R. and Whitelaw,E. (2004). Retrotransposon-derived elements in the mammalian genome: a potential source of disease. J. Inherit. Metab Dis. *27*, 319-330.
- Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J., and Gejman,P.V. (2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. *12*, 205-216.
- Duan,S., Gao,R., Xing,Q., Du,J., Liu,Z., Chen,Q., Wang,H., Feng,G., and He,L. (2005). A family-based association study of schizophrenia with polymorphisms at three candidate genes. Neurosci. Lett. *379*, 32-36.
- Dubchak,I., Brudno,M., Loots,G.G., Pachter,L., Mayor,C., Rubin,E.M., and Frazer,K.A. (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. *10*, 1304-1306.
- Dubchak,I. and Frazer,K. (2003). Multi-species sequence comparison: the next frontier in genome annotation. Genome Biol. *4*, 122.
- Dubertret,C., Gorwood,P., Gouya,L., Deybach,J.C., and Ades,J. (2001). Association and excess of transmission of a DRD2 haplotype in a sample of French schizophrenic patients. Schizophr. Res. *49*, 203-212.
- Dubois,P. (1999). MySQL. (Indianapolis: New Riders Publishing).
- Duchateau,P.N., Pullinger,C.R., Cho,M.H., Eng,C., and Kane,J.P. (2001). Apolipoprotein L gene family: tissue-specific expression, splicing, promoter regions; discovery of a new gene. J. Lipid Res. *42*, 620-630.
- Duggal,P., An,P., Beaty,T.H., Strathdee,S.A., Farzadegan,H., Markham,R.B., Johnson,L., O'Brien,S.J., Vlahov,D., and Winkler,C.A. (2003). Genetic influence of CXCR6 chemokine receptor alleles on PCP-mediated AIDS progression among African Americans. Genes Immun. *4*, 245-250.
- Duret,L. and Bucher,P. (1997). Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. *7*, 399-406.
- Dvir,A. (2002). Promoter escape by RNA polymerase II. Biochim. Biophys. Acta *1577*, 208-223.

- Ebihara,M., Ohba,H., Ohno,S.I., and Yoshikawa,T. (2002). Genomic organization and promoter analysis of the human nicotinic acetylcholine receptor alpha6 subunit (CHNRA6) gene: Alu and other elements direct transcriptional repression. Gene *298*, 101-108.
- Egan,M.F., Goldberg,T.E., Kolachana,B.S., Callicott,J.H., Mazzanti,C.M., Straub,R.E., Goldman,D., and Weinberger,D.R. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. Proc. Natl. Acad. Sci. U. S. A *98*, 6917-6922.
- Egan,M.F., Straub,R.E., Goldberg,T.E., Yakub,I., Callicott,J.H., Hariri,A.R., Mattay,V.S., Bertolino,A., Hyde,T.M., Shannon-Weickert,C., Akil,M., Crook,J., Vakkalanka,R.K., Balkissoon,R., Gibbs,R.A., Kleinman,J.E., and Weinberger,D.R. (2004). Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. Proc. Natl. Acad. Sci. U. S. A *101*, 12604-12609.
- Ekelund,J., Hennah,W., Hiekkalinna,T., Parker,A., Meyer,J., Lonnqvist,J., and Peltonen,L. (2004). Replication of 1q42 linkage in Finnish schizophrenia pedigrees. Mol. Psychiatry *9*, 1037-1041.
- Ekelund,J., Hovatta,I., Parker,A., Paunio,T., Varilo,T., Martin,R., Suhonen,J., Ellonen,P., Chan,G., Sinsheimer,J.S., Sobel,E., Juvonen,H., Arajarvi,R., Partonen,T., Suvisaari,J., Lonnqvist,J., Meyer,J., and Peltonen,L. (2001). Chromosome 1 loci in Finnish schizophrenia families. Hum. Mol. Genet. *10*, 1611-1617.
- el Husseini,A. and Bredt,D.S. (2002). Protein palmitoylation: a regulator of neuronal development and function. Nat. Rev. Neurosci. *3*, 791-802.
- Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W., and Chiaromonte,F. (2003). Distinguishing regulatory DNA from neutral sites. Genome Res. *13*, 64-72.
- Elnitski,L., Riemer,C., Petrykowska,H., Florea,L., Schwartz,S., Miller,W., and Hardison,R. (2002). PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. Genomics *80*, 681-690.
- Elvidge,G., Jones,I., McCandless,F., Asherson,P., Owen,M.J., and Craddock,N. (2001). Allelic variation of a BalI polymorphism in the DRD3 gene does not influence susceptibility to bipolar disorder: results of analysis and meta-analysis. Am. J. Med. Genet. *105*, 307-311.
- Emamian,E.S., Hall,D., Birnbaum,M.J., Karayiorgou,M., and Gogos,J.A. (2004). Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia. Nat. Genet. *36*, 131-137.
- Emilien,G., Maloteaux,J.M., Geurts,M., Hoogenberg,K., and Cragg,S. (1999). Dopamine receptors--physiological understanding to therapeutic intervention potential. Pharmacol. Ther. *84*, 133-156.
- Enard,W., Khaitovich,P., Klose,J., Zollner,S., Heissig,F., Giavalisco,P., Nieselt-Struwe,K., Muchmore,E., Varki,A., Ravid,R., Doxiadis,G.M., Bontrop,R.E., and Paabo,S. (2002). Intra- and interspecific variation in primate gene expression patterns. Science *296*, 340-343.
- Escamilla,M.A., McInnes,L.A., Spesny,M., Reus,V.I., Service,S.K., Shimayoshi,N., Tyler,D.J., Silva,S., Molina,J., Gallegos,A., Meza,L., Cruz,M.L., Batki,S., Vinogradov,S., Neylan,T., Nguyen,J.B., Fournier,E., Araya,C., Barondes,S.H., Leon,P., Sandkuijl,L.A., and Freimer,N.B. (1999). Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. Am. J. Hum. Genet. *64*, 1670-1678.
- Evans,K.L., Le Hellard,S., Morris,S.W., Lawson,D., Whitton,C., Semple,C.A., Fantes,J.A., Torrance,H.S., Malloy,M.P., Maule,J.C., Humphray,S.J., Ross,M.T., Bentley,D.R., Muir,W.J., Blackwood,D.H., and Porteous,D.J. (2001). A 6.9-Mb high-resolution BAC/PAC contig of human 4p15.3-p16.1, a candidate region for bipolar affective disorder. Genomics *71*, 315-323.
- Ewald,H., Degn,B., Mors,O., and Kruse,T.A. (1998). Support for the possible locus on chromosome 4p16 for bipolar affective disorder. Mol. Psychiatry *3*, 442-448.
- Ewald,H., Flint,T., Kruse,T.A., and Mors,O. (2002). A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24.3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13.3. Mol. Psychiatry *7*, 734-744.

- Ewens,W.J. and Spielman,R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. *57*, 455-464.
- Faber,E.S. and Sah,P. (2003). Calcium-activated potassium channels: multiple contributions to neuronal function. Neuroscientist *9*, 181-194.
- Fabian-Fine,R., Skehel,P., Errington,M.L., Davies,H.A., Sher,E., Stewart,M.G., and Fine,A. (2001). Ultrastructural distribution of the alpha7 nicotinic acetylcholine receptor subunit in rat hippocampus. J. Neurosci. *21*, 7993-8003.
- Falls,D.L. (2003). Neuregulins: functions, forms, and signaling strategies. Exp. Cell Res. *284*, 14-30.
- Falret (1854). Mémoire sur la folie circulaire, forme de la maladie mentale caractérisée par la reproduction successive et régulière de l'état maniaque, de l'état mélancolique, et d'un intervalle lucide plus ou moins prolongé. Bulletin de l'académie impériale de médecine *19*, 382-400.
- Fan,J., Fossella,J., Sommer,T., Wu,Y., and Posner,M.I. (2003a). Mapping the genetic variation of executive attention onto brain activity. Proc. Natl. Acad. Sci. U. S. A *100*, 7406-7411.
- Fan,J.B., Ma,J., Zhang,C.S., Tang,J.X., Gu,N.F., Feng,G.Y., St Clair,D., and He,L. (2003b). A family-based association study of T1945C polymorphism in the proline dehydrogenase gene and schizophrenia in the Chinese population. Neurosci. Lett. *338*, 252-254.
- Fanous,A.H., Neale,M.C., Straub,R.E., Webb,B.T., O'Neill,A.F., Walsh,D., and Kendler,K.S. (2004). Clinical features of psychotic disorders and polymorphisms in HT2A, DRD2, DRD4, SLC6A3 (DAT1), and BDNF: a family based association study. Am. J. Med. Genet. B Neuropsychiatr. Genet. *125*, 69-78.
- Fay,J.C., McCullough,H.L., Sniegowski,P.D., and Eisen,M.B. (2004). Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol. *5*, R26.
- Ferrigno,O., Virolle,T., Djabari,Z., Ortonne,J.P., White,R.J., and Aberdam,D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nat. Genet. *28*, 77-81.
- Fessele,S., Maier,H., Zischek,C., Nelson,P.J., and Werner,T. (2002). Regulatory context is a crucial part of gene function. Trends Genet. *18*, 60-63.
- Figueroa,K.P., Chan,P., Schols,L., Tanner,C., Riess,O., Perlman,S.L., Geschwind,D.H., and Pulst,S.M. (2001). Association of moderate polyglutamine tract expansions in the slow calcium-activated potassium channel type 3 with ataxia. Arch. Neurol. *58*, 1649-1653.
- Fiskerstrand,C.E., Lovejoy,E.A., and Quinn,J.P. (1999). An intronic polymorphic domain often associated with susceptibility to affective disorders has allele dependent differential enhancer activity in embryonic stem cells. FEBS Lett. *458*, 171-174.
- Flicek,P., Keibler,E., Hu,P., Korf,I., and Brent,M.R. (2003). Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res. *13*, 46-54.
- Forget,D., Robert,F., Grondin,G., Burton,Z.F., Greenblatt,J., and Coulombe,B. (1997). RAP74 induces promoter contacts by RNA polymerase II upstream and downstream of a DNA bend centered on the TATA box. Proc. Natl. Acad. Sci. U. S. A *94*, 7150-7155.
- Fortini,M.E. and Artavanis-Tsakonas,S. (1993). Notch: neurogenesis is only part of the picture. Cell *75*, 1245-1247.
- Fossella,J., Sommer,T., Fan,J., Wu,Y., Swanson,J.M., Pfaff,D.W., and Posner,M.I. (2002). Assessing the molecular genetics of attention networks. BMC. Neurosci. *3*, 14.
- Franchini,L., Serretti,A., Gasperini,M., and Smeraldi,E. (1998). Familial concordance of fluvoxamine response as a tool for differentiating mood disorder pedigrees. J. Psychiatr. Res. *32*, 255-259.
- Frank,C.G., Grubenmann,C.E., Eyaid,W., Berger,E.G., Aebi,M., and Hennet,T. (2004). Identification and functional analysis of a defect in the human ALG9 gene: definition of congenital disorder of glycosylation type IL. Am. J. Hum. Genet. *75*, 146-150.

- Frazer,K.A., Elnitski,L., Church,D.M., Dubchak,I., and Hardison,R.C. (2003). Cross-species sequence comparisons: a review of methods and available resources. Genome Res. *13*, 1-12.
- Frazier,C.J., Strowbridge,B.W., and Papke,R.L. (2003). Nicotinic receptors on local circuit neurons in dentate gyrus: a potential role in regulation of granule cell excitability. J. Neurophysiol. *89*, 3018-3028.
- Freedman,R., Leonard,S., Gault,J.M., Hopkins,J., Cloninger,C.R., Kaufmann,C.A., Tsuang,M.T., Farone,S.V., Malaspina,D., Svrakic,D.M., Sanders,A., and Gejman,P. (2001). Linkage disequilibrium for schizophrenia at the chromosome 15q13-14 locus of the alpha7-nicotinic acetylcholine receptor subunit gene (CHRNA7). Am. J. Med. Genet. *105*, 20-22.
- Fremeau,R.T., Jr., Caron,M.G., and Blakely,R.D. (1992). Molecular cloning and expression of a high affinity L-proline transporter expressed in putative glutamatergic pathways of rat brain. Neuron *8*, 915-926.
- Frisch,M., Frech,K., Klingenhoff,A., Cartharius,K., Liebich,I., and Werner,T. (2002). In silico prediction of scaffold/matrix attachment regions in large genomic sequences. Genome Res. *12*, 349-354.
- Frith,M.C., Li,M.C., and Weng,Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. *31*, 3666-3668.
- Fujii,Y., Shibata,H., Kikuta,R., Makino,C., Tani,A., Hirata,N., Shibata,A., Ninomiya,H., Tashiro,N., and Fukumaki,Y. (2003). Positive associations of polymorphisms in the metabotropic glutamate receptor type 3 gene (GRM3) with schizophrenia. Psychiatr. Genet. *13*, 71-76.
- Fukuda,T., Yamamoto,I., Nishida,Y., Zhou,Q., Ohno,M., Takada,K., and Azuma,J. (1999). Effect of the CYP2D6*10 genotype on venlafaxine pharmacokinetics in healthy adult volunteers. Br. J. Clin. Pharmacol. *47*, 450-453.
- Funakoshi,A., Miyasaka,K., Matsumoto,H., Yamamori,S., Takiguchi,S., Kataoka,K., Takata,Y., Matsusue,K., Kono,A., and Shimokata,H. (2000). Gene structure of human cholecystokinin (CCK) type-A receptor: body fat content is related to CCK type-A receptor gene promoter polymorphism. FEBS Lett. *466*, 264-266.
- Furano,A.V. (2000). The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog. Nucleic Acid Res. Mol. Biol. *64*, 255-294.
- Furlong,R.A., Ho,L., Rubinsztein,J.S., Walsh,C., Paykel,E.S., and Rubinsztein,D.C. (1999a). Analysis of the monoamine oxidase A (MAOA) gene in bipolar affective disorder by association studies, meta-analyses, and sequencing of the promoter. Am. J. Med. Genet. *88*, 398-406.
- Furlong,R.A., Rubinsztein,J.S., Ho,L., Walsh,C., Coleman,T.A., Muir,W.J., Paykel,E.S., Blackwood,D.H., and Rubinsztein,D.C. (1999b). Analysis and metaanalysis of two polymorphisms within the tyrosine hydroxylase gene in bipolar and unipolar affective disorders. Am. J. Med. Genet. *88*, 88-94.
- Gabellini,D., Tupler,R., and Green,M.R. (2003). Transcriptional derepression as a cause of genetic diseases. Curr. Opin. Genet. Dev. *13*, 239-245.
- Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M., Liu-Cordero,S.N., Rotimi,C., Adeyemo,A., Cooper,R., Ward,R., Lander,E.S., Daly,M.J., and Altshuler,D. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.
- Gailus-Durner,V., Scherf,M., and Werner,T. (2001). Experimental data of a single promoter can be used for in silico detection of genes with related regulation in the absence of sequence similarity. Mamm. Genome *12*, 67-72.
- Galant,R. and Carroll,S.B. (2002). Evolution of a transcriptional repression domain in an insect Hox protein. Nature *415*, 910-913.
- Gargus,J.J., Fantino,E., and Gutman,G.A. (1998). A piece in the puzzle: an ion channel candidate gene for schizophrenia. Mol. Med. Today *4*, 518-524.
- Gasperoni,T.L., Ekelund,J., Huttunen,M., Palmer,C.G., Tuulio-Henriksson,A., Lonnqvist,J., Kaprio,J., Peltonen,L., and Cannon,T.D. (2003). Genetic linkage and association between

chromosome 1q and working memory function in schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *116*, 8-16.

- Gecz,J., Barnett,S., Liu,J., Hollway,G., Donnelly,A., Eyre,H., Eshkevari,H.S., Baltazar,R., Grunn,A., Nagaraja,R., Gilliam,C., Peltonen,L., Sutherland,G.R., Baron,M., and Mulley,J.C. (1999). Characterization of the human glutamate receptor subunit 3 gene (GRIA3), a candidate for bipolar disorder and nonspecific X-linked mental retardation. Genomics *62*, 356-368.
- Gelder,M., Mayou,R., and Geffers,R. (2000). Psychiatry. (Oxford: Oxford University Press).
- George,T.P., Verrico,C.D., Picciotto,M.R., and Roth,R.H. (2000). Nicotinic modulation of mesoprefrontal dopamine neurons: pharmacologic and neuroanatomic characterization. J. Pharmacol. Exp. Ther. *295*, 58-66.
- Geracioti,T.D. and Liddle,R.A. (1988). Impaired cholecystokinin secretion in bulimia nervosa. N. Engl. J. Med. *319*, 683-688.
- Gibbs,J., Young,R.C., and Smith,G.P. (1973). Cholecystokinin elicits satiety in rats with open gastric fistulas. Nature *245*, 323-325.
- Gibbs,R.A., Weinstock,G.M., Metzker,M.L., Muzny,D.M., Sodergren,E.J., Scherer,S., Scott,G., Steffen,D., Worley,K.C., Burch,P.E., Okwuonu,G., Hines,S., Lewis,L., DeRamo,C., Delgado,O., Dugan-Rocha,S., Miner,G., Morgan,M., Hawes,A., Gill,R., Celera, Holt,R.A., Adams,M.D., Amanatides,P.G., Baden-Tillson,H., Barnstead,M., Chin,S., Evans,C.A., Ferriera,S., Fosler,C., Glodek,A., Gu,Z., Jennings,D., Kraft,C.L., Nguyen,T., Pfannkoch,C.M., Sitter,C., Sutton,G.G., Venter,J.C., Woodage,T., Smith,D., Lee,H.M., Gustafson,E., Cahill,P., Kana,A., Doucette-Stamm,L., Weinstock,K., Fechtel,K., Weiss,R.B., Dunn,D.M., Green,E.D., Blakesley,R.W., Bouffard,G.G., De Jong,P.J., Osoegawa,K., Zhu,B., Marra,M., Schein,J., Bosdet,I., Fjell,C., Jones,S., Krzywinski,M., Mathewson,C., Siddiqui,A., Wye,N., McPherson,J., Zhao,S., Fraser,C.M., Shetty,J., Shatsman,S., Geer,K., Chen,Y., Abramzon,S., Nierman,W.C., Havlak,P.H., Chen,R., Durbin,K.J., Egan,A., Ren,Y., Song,X.Z., Li,B., Liu,Y., Qin,X., Cawley,S., Worley,K.C., Cooney,A.J., D'Souza,L.M., Martin,K., Wu,J.Q., Gonzalez-Garay,M.L., Jackson,A.R., Kalafus,K.J., McLeod,M.P., Milosavljevic,A., Virk,D., Volkov,A., Wheeler,D.A., Zhang,Z., Bailey,J.A., Eichler,E.E., Tuzun,E., Birney,E., Mongin,E., Ureta-Vidal,A., Woodwark,C., Zdobnov,E., Bork,P., Suyama,M., Torrents,D., Alexandersson,M., Trask,B.J., Young,J.M., Huang,H., Wang,H., Xing,H., Daniels,S., Gietzen,D., Schmidt,J., Stevens,K., Vitt,U., Wingrove,J., Camara,F., Mar,A.M., Abril,J.F., Guigo,R., Smit,A., Dubchak,I., Rubin,E.M., Couronne,O., Poliakov,A., Hubner,N., Ganten,D., Goesele,C., Hummel,O., Kreitler,T., Lee,Y.A., Monti,J., Schulz,H., Zimdahl,H., Himmelbauer,H., Lehrach,H., Jacob,H.J., Bromberg,S., Gullings-Handley,J., Jensen-Seaman,M.I., Kwitek,A.E., Lazar,J., Pasko,D., Tonellato,P.J., Twigger,S., Ponting,C.P., Duarte,J.M., Rice,S., Goodstadt,L., Beatson,S.A., Emes,R.D., Winter,E.E., Webber,C., Brandt,P., Nyakatura,G., Adetobi,M., Chiaromonte,F., Elnitski,L., Eswara,P., Hardison,R.C., Hou,M., Kolbe,D., Makova,K., Miller,W., Nekrutenko,A., Riemer,C., Schwartz,S., Taylor,J., Yang,S., Zhang,Y., Lindpaintner,K., Andrews,T.D., Caccamo,M., Clamp,M., Clarke,L., Curwen,V., Durbin,R., Eyras,E., Searle,S.M., Cooper,G.M., Batzoglou,S., Brudno,M., Sidow,A., Stone,E.A., Venter,J.C., Payseur,B.A., Bourque,G., Lopez-Otin,C., Puente,X.S., Chakrabarti,K., Chatterji,S., Dewey,C., Pachter,L., Bray,N., Yap,V.B., Caspi,A., Tesler,G., Pevzner,P.A., Haussler,D., Roskin,K.M., Baertsch,R., Clawson,H., Furey,T.S., Hinrichs,A.S., Karolchik,D., Kent,W.J., Rosenbloom,K.R., Trumbower,H., Weirauch,M., Cooper,D.N., Stenson,P.D., Ma,B., Brent,M., Arumugam,M., Shteynberg,D., Copley,R.R., Taylor,M.S., Riethman,H., Mudunuri,U., Peterson,J., Guyer,M., Felsenfeld,A., Old,S., Mockrin,S., and Collins,F. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.
- Ginns,E.I., St Jean,P., Philibert,R.A., Galdzicka,M., Damschroder-Williams,P., Thiel,B., Long,R.T., Ingraham,L.J., Dalwaldi,H., Murray,M.A., Ehlert,M., Paul,S., Remortel,B.G., Patel,A.P., Anderson,M.C., Shaio,C., Lau,E., Dymarskaia,I., Martin,B.M., Stubblefield,B., Falls,K.M., Carulli,J.P., Keith,T.P., Fann,C.S., Lacy,L.G., Allen,C.R., Hostetter,A.M., Elston,R.C., Schork,N.J., Egeland,J.A., and Paul,S.M. (1998). A genome-wide search for chromosomal loci linked to mental health wellness in relatives at high risk for bipolar

affective disorder among the Old Order Amish. Proc. Natl. Acad. Sci. U. S. A *95*, 15531-15536.

- Glatt,S.J., Faraone,S.V., and Tsuang,M.T. (2003a). Association between a functional catechol O-methyltransferase gene polymorphism and schizophrenia: meta-analysis of case-control and family-based studies. Am. J. Psychiatry *160*, 469-476.
- Glatt,S.J., Faraone,S.V., and Tsuang,M.T. (2003b). Meta-analysis identifies an association between the dopamine D2 receptor gene and schizophrenia. Mol. Psychiatry *8*, 911-915.
- Glatt,S.J., Faraone,S.V., and Tsuang,M.T. (2004). DRD2 -141C insertion/deletion polymorphism is not associated with schizophrenia: results of a meta-analysis. Am. J. Med. Genet. B Neuropsychiatr. Genet. *128*, 21-23.
- Glatt,S.J., Wang,R.S., Yeh,Y.C., Tsuang,M.T., and Faraone,S.V. (2005). Five NOTCH4 polymorphisms show weak evidence for association with schizophrenia: evidence from meta-analyses. Schizophr. Res. *73*, 281-290.
- Gogos,J.A., Santha,M., Takacs,Z., Beck,K.D., Luine,V., Lucas,L.R., Nadler,J.V., and Karayiorgou,M. (1999). The gene encoding proline dehydrogenase modulates sensorimotor gating in mice. Nat. Genet. *21*, 434-439.
- Goldman-Rakic,P.S. (1999). The physiological approach: functional architecture of working memory and disordered cognition in schizophrenia. Biol. Psychiatry *46*, 650-661.
- Goldman-Rakic,P.S. and Selemon,L.D. (1997). Functional and anatomical aspects of prefrontal pathology in schizophrenia. Schizophr. Bull. *23*, 437-458.
- Goldstein,D.B., Ahmadi,K.R., Weale,M.E., and Wood,N.W. (2003). Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. Trends Genet. *19*, 615-622.
- Golimbet,V.E., Aksenova,M.G., Nosikov,V.V., Orlova,V.A., and Kaleda,V.G. (2003). Analysis of the linkage of the Taq1A and Taq1B loci of the dopamine D2 receptor gene with schizophrenia in patients and their siblings. Neurosci. Behav. Physiol *33*, 223-225.
- Gomez-Casero,E., Perez,d.C., I, Saiz-Ruiz,J., Llinares,C., and Fernandez-Piqueras,J. (1996). No association between particular DRD3 and DAT gene polymorphisms and manic-depressive illness in a Spanish sample. Psychiatr. Genet. *6*, 209-212.
- Gompel,N., Prud'homme,B., Wittkopp,P.J., Kassner,V.A., and Carroll,S.B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature *433*, 481-487.
- Gonen,D., Veenstra-VanderWeele,J., Yang,Z., Leventhal,B., and Cook,E.H., Jr. (1999). High throughput fluorescent CE-SSCP SNP genotyping. Mol. Psychiatry *4*, 339-343.
- Goodman,N. (2002). The serotonergic system and mysticism: could LSD and the nondrug-induced mystical experience share common neural mechanisms? J. Psychoactive Drugs *34*, 263-272.
- Gottgens,B., Barton,L.M., Chapman,M.A., Sinclair,A.M., Knudsen,B., Grafham,D., Gilbert,J.G., Rogers,J., Bentley,D.R., and Green,A.R. (2002). Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. Genome Res. *12*, 749-759.
- Gottgens,B., Barton,L.M., Gilbert,J.G., Bench,A.J., Sanchez,M.J., Bahn,S., Mistry,S., Grafham,D., McMurray,A., Vaudin,M., Amaya,E., Bentley,D.R., Green,A.R., and Sinclair,A.M. (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. Nat. Biotechnol. *18*, 181-186.
- Grandy,D.K., Zhang,Y.A., Bouvier,C., Zhou,Q.Y., Johnson,R.A., Allen,L., Buck,K., Bunzow,J.R., Salon,J., and Civelli,O. (1991). Multiple human D5 dopamine receptor genes: a functional receptor and two pseudogenes. Proc. Natl. Acad. Sci. U. S. A *88*, 9175-9179.
- Graves,J.A. and Westerman,M. (2002). Marsupial genetics and genomics. Trends Genet. *18*, 517-521.
- Green, M.F. (2001). Schizophrenia revealed: from neurons to social interactions. (New York: Norton W.W.).

- Greenberg,P.E., Kessler,R.C., Birnbaum,H.G., Leong,S.A., Lowe,S.W., Berglund,P.A., and Corey-Lisle,P.K. (2003). The economic burden of depression in the United States: how did it change between 1990 and 2000? J. Clin. Psychiatry *64*, 1465-1475.
- Greene,E., Entezam,A., Kumari,D., and Usdin,K. (2005). Ancient repeated DNA elements and the regulation of the human frataxin promoter. Genomics *85*, 221-230.
- Greenwood,T.A., Alexander,M., Keck,P.E., McElroy,S., Sadovnick,A.D., Remick,R.A., and Kelsoe,J.R. (2001). Evidence for linkage disequilibrium between the dopamine transporter and bipolar disorder. Am. J. Med. Genet. *105*, 145-151.
- Greenwood,T.A. and Kelsoe,J.R. (2003). Promoter and intronic variants affect the transcriptional regulation of the human dopamine transporter gene. Genomics *82*, 511-520.
- Gretarsdottir,S., Thorleifsson,G., Reynisdottir,S.T., Manolescu,A., Jonsdottir,S., Jonsdottir,T., Gudmundsdottir,T., Bjarnadottir,S.M., Einarsson,O.B., Gudjonsdottir,H.M., Hawkins,M., Gudmundsson,G., Gudmundsdottir,H., Andrason,H., Gudmundsdottir,A.S., Sigurdardottir,M., Chou,T.T., Nahmias,J., Goss,S., Sveinbjornsdottir,S., Valdimarsson,E.M., Jakobsson,F., Agnarsson,U., Gudnason,V., Thorgeirsson,G., Fingerle,J., Gurney,M., Gudbjartsson,D., Frigge,M.L., Kong,A., Stefansson,K., and Gulcher,J.R. (2003). The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat. Genet. *35*, 131-138.
- Grillet,N., Dubreuil,V., Dufour,H.D., and Brunet,J.F. (2003). Dynamic expression of RGS4 in the developing nervous system and regulation by the neural type-specific transcription factor Phox2b. J. Neurosci. *23*, 10613-10621.
- Grunstein,M. (1998). Yeast heterochromatin: regulation of its assembly and inheritance by histones. Cell *93*, 325-328.
- Guillin,O., Diaz,J., Carroll,P., Griffon,N., Schwartz,J.C., and Sokoloff,P. (2001). BDNF controls dopamine D3 receptor expression and triggers behavioural sensitization. Nature *411*, 86-89.
- Gumucio,D.L., Shelton,D.A., Zhu,W., Millinoff,D., Gray,T., Bock,J.H., Slightom,J.L., and Goodman,M. (1996). Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. Mol. Phylogenet. Evol. *5*, 18-32.
- Guo,J., Wenk,M.R., Pellegrini,L., Onofri,F., Benfenati,F., and De Camilli,P. (2003). Phosphatidylinositol 4-kinase type IIalpha is responsible for the phosphatidylinositol 4-kinase activity associated with synaptic vesicles. Proc. Natl. Acad. Sci. U. S. A *100*, 3995-4000.
- Gurling,H.M., Kalsi,G., Brynjolfson,J., Sigmundsson,T., Sherrington,R., Mankoo,B.S., Read,T., Murphy,P., Blaveri,E., McQuillin,A., Petursson,H., and Curtis,D. (2001). Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. Am. J. Hum. Genet. *68*, 661-673.
- Gusella,J.F., Wexler,N.S., Conneally,P.M., Naylor,S.L., Anderson,M.A., Tanzi,R.E., Watkins,P.C., Ottina,K., Wallace,M.R., and Sakaguchi,A.Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.
- Guzey,C. and Spigset,O. (2004). Genotyping as a tool to predict adverse drug reactions. Curr. Top. Med. Chem. *4*, 1411-1421.
- Hall,J.M., Lee,M.K., Newman,B., Morrow,J.E., Anderson,L.A., Huey,B., and King,M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684-1689.
- Hallmayer,J. (2004). Getting our AKT together in schizophrenia? Nat. Genet. *36*, 115-116.
- Hambor,J.E., Mennone,J., Coon,M.E., Hanke,J.H., and Kavathas,P. (1993). Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. Mol. Cell Biol. *13*, 7056-7070.
- Hamilton,S.P., Slager,S.L., Helleby,L., Heiman,G.A., Klein,D.F., Hodge,S.E., Weissman,M.M., Fyer,A.J., and Knowles,J.A. (2001). No association or linkage between

polymorphisms in the genes encoding cholecystokinin and the cholecystokinin B receptor and panic disorder. Mol. Psychiatry *6*, 59-65.

- Hansen,T.O., Rehfeld,J.F., and Nielsen,F.C. (2004). GSK-3beta reduces cAMP-induced cholecystokinin gene expression by inhibiting CREB binding. Neuroreport *15*, 841-845.
- Hansson,E.M., Lendahl,U., and Chapman,G. (2004). Notch signaling in development and disease. Semin. Cancer Biol. *14*, 320-328.
- Harada,S., Okubo,T., Tsutsumi,M., Takase,S., and Muramatsu,T. (1998). A new genetic variant in the Sp1 binding cis-element of cholecystokinin gene promoter region and relationship to alcoholism. Alcohol Clin. Exp. Res. *22*, 93S-96S.
- Hardison,R., Krane,D., Vandenbergh,D., Cheng,J.F., Mansberger,J., Taddie,J., Schwartz,S., Huang,X.Q., and Miller,W. (1991). Sequence and comparative analysis of the rabbit alpha-like globin gene cluster reveals a rapid mode of evolution in a G + C-rich region of mammalian genomes. J. Mol. Biol. *222*, 233-249.
- Hardison,R.C., Chui,D.H., Giardine,B., Riemer,C., Patrinos,G.P., Anagnou,N., Miller,W., and Wajcman,H. (2002). HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum. Mutat. *19*, 225-233.
- Hardison,R.C., Oeltjen,J., and Miller,W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. *7*, 959-966.
- Hariri,A.R., Mattay,V.S., Tessitore,A., Kolachana,B., Fera,F., Goldman,D., Egan,M.F., and Weinberger,D.R. (2002). Serotonin transporter genetic variation and the response of the human amygdala. Science *297*, 400-403.
- Harrison,P.J. (1997). Schizophrenia: a disorder of neurodevelopment? Curr. Opin. Neurobiol. *7*, 285-289.
- Harrison,P.J. and Weinberger,D.R. (2005). Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol. Psychiatry *10*, 40-68.
- Harwood,A.J. and Agam,G. (2003). Search for a common mechanism of mood stabilizers. Biochem. Pharmacol. *66*, 179-189.
- Hashimoto,R., Straub,R.E., Weickert,C.S., Hyde,T.M., Kleinman,J.E., and Weinberger,D.R. (2004). Expression analysis of neuregulin-1 in the dorsolateral prefrontal cortex in schizophrenia. Mol. Psychiatry *9*, 299-307.
- Hattori,E., Ebihara,M., Yamada,K., Ohba,H., Shibuya,H., and Yoshikawa,T. (2001). Identification of a compound short tandem repeat stretch in the 5'-upstream region of the cholecystokinin gene, and its association with panic disorder but not with schizophrenia. Mol. Psychiatry *6*, 465-470.
- Hattori,E., Liu,C., Badner,J.A., Bonner,T.I., Christian,S.L., Maheshwari,M., Detera-Wadleigh,S.D., Gibbs,R.A., and Gershon,E.S. (2003). Polymorphisms at the G72/G30 gene locus, on 13q33, are associated with bipolar disorder in two independent pedigree series. Am. J. Hum. Genet. *72*, 1131-1140.
- Hattori,E., Yamada,K., Ebihara,M., Toyota,T., Nankai,M., Shibuya,H., and Yoshikawa,T. (2002). Association study of the short tandem repeat in the 5' upstream region of the cholecystokinin gene with mood disorders in the Japanese population. Am. J. Med. Genet. *114*, 523-526.
- Hawi,Z., Lowe,N., Kirley,A., Gruenhage,F., Nothen,M., Greenwood,T., Kelsoe,J., Fitzgerald,M., and Gill,M. (2003). Linkage disequilibrium mapping at DAT1, DRD5 and DBH narrows the search for ADHD susceptibility alleles at these loci. Mol. Psychiatry *8*, 299-308.
- Hawi,Z., Mynett-Johnson,L., Murphy,V., Straub,R.E., Kendler,K.S., Walsh,D., McKeon,P., and Gill,M. (1999). No evidence to support the association of the potassium channel gene hSKCa3 CAG repeat with schizophrenia or bipolar disorder in the Irish population. Mol. Psychiatry *4*, 488-491.
- Heilmeyer,L.M., Vereb,G., Jr., Vereb,G., Kakuk,A., and Szivak,I. (2003). Mammalian phosphatidylinositol 4-kinases. IUBMB. Life *55*, 59-65.

- Heils,A., Teufel,A., Petri,S., Stober,G., Riederer,P., Bengel,D., and Lesch,K.P. (1996). Allelic variation of human serotonin transporter gene expression. J. Neurochem. *66*, 2621-2624.
- Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L., and Kolchanov,N.A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res. *26*, 362-367.
- Heinz,A., Jones,D.W., Mazzanti,C., Goldman,D., Ragan,P., Hommer,D., Linnoila,M., and Weinberger,D.R. (2000). A relationship between serotonin transporter genotype and in vivo protein expression and alcohol neurotoxicity. Biol. Psychiatry *47*, 643-649.
- Helgadottir,A., Manolescu,A., Thorleifsson,G., Gretarsdottir,S., Jonsdottir,H., Thorsteinsdottir,U., Samani,N.J., Gudmundsson,G., Grant,S.F., Thorgeirsson,G., Sveinbjornsdottir,S., Valdimarsson,E.M., Matthiasson,S.E., Johannsson,H., Gudmundsdottir,O., Gurney,M.E., Sainz,J., Thorhallsdottir,M., Andresdottir,M., Frigge,M.L., Topol,E.J., Kong,A., Gudnason,V., Hakonarson,H., Gulcher,J.R., and Stefansson,K. (2004). The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat. Genet. *36*, 233-239.
- Hennah,W., Varilo,T., Kestila,M., Paunio,T., Arajarvi,R., Haukka,J., Parker,A., Martin,R., Levitzky,S., Partonen,T., Meyer,J., Lonnqvist,J., Peltonen,L., and Ekelund,J. (2003). Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. Hum. Mol. Genet. *12*, 3151-3159.
- Heresco-Levy,U., Javitt,D.C., Ermilov,M., Mordel,C., Horowitz,A., and Kelly,D. (1996). Double-blind, placebo-controlled, crossover trial of glycine adjuvant therapy for treatment-resistant schizophrenia. Br. J. Psychiatry *169*, 610-617.
- Hillebrand,J.J., de Wied,D., and Adan,R.A. (2002). Neuropeptides, food intake and body weight regulation: a hypothalamic focus. Peptides *23*, 2283-2306.
- Hillier,L.W., Miller,W., Birney,E., Warren,W., Hardison,R.C., Ponting,C.P., Bork,P., Burt,D.W., Groenen,M.A., Delany,M.E., Dodgson,J.B., Chinwalla,A.T., Cliften,P.F., Clifton,S.W., Delehaunty,K.D., Fronick,C., Fulton,R.S., Graves,T.A., Kremitzki,C., Layman,D., Magrini,V., McPherson,J.D., Miner,T.L., Minx,P., Nash,W.E., Nhan,M.N., Nelson,J.O., Oddy,L.G., Pohl,C.S., Randall-Maher,J., Smith,S.M., Wallis,J.W., Yang,S.P., Romanov,M.N., Rondelli,C.M., Paton,B., Smith,J., Morrice,D., Daniels,L., Tempest,H.G., Robertson,L., Masabanda,J.S., Griffin,D.K., Vignal,A., Fillon,V., Jacobbson,L., Kerje,S., Andersson,L., Crooijmans,R.P., Aerts,J., van der Poel,J.J., Ellegren,H., Caldwell,R.B., Hubbard,S.J., Grafham,D.V., Kierzek,A.M., McLaren,S.R., Overton,I.M., Arakawa,H., Beattie,K.J., Bezzubov,Y., Boardman,P.E., Bonfield,J.K., Croning,M.D., Davies,R.M., Francis,M.D., Humphray,S.J., Scott,C.E., Taylor,R.G., Tickle,C., Brown,W.R., Rogers,J., Buerstedde,J.M., Wilson,S.A., Stubbs,L., Ovcharenko,I., Gordon,L., Lucas,S., Miller,M.M., Inoko,H., Shiina,T., Kaufman,J., Salomonsen,J., Skjoedt,K., Wong,G.K., Wang,J., Liu,B., Wang,J., Yu,J., Yang,H., Nefedov,M., Koriabine,M., Dejong,P.J., Goodstadt,L., Webber,C., Dickens,N.J., Letunic,I., Suyama,M., Torrents,D., von Mering,C., Zdobnov,E.M., Makova,K., Nekrutenko,A., Elnitski,L., Eswara,P., King,D.C., Yang,S., Tyekucheva,S., Radakrishnan,A., Harris,R.S., Chiaromonte,F., Taylor,J., He,J., Rijnkels,M., Griffiths-Jones,S., Ureta-Vidal,A., Hoffman,M.M., Severin,J., Searle,S.M., Law,A.S., Speed,D., Waddington,D., Cheng,Z., Tuzun,E., Eichler,E., Bao,Z., Flicek,P., Shteynberg,D.D., Brent,M.R., Bye,J.M., Huckle,E.J., Chatterji,S., Dewey,C., Pachter,L., Kouranov,A., Mourelatos,Z., Hatzigeorgiou,A.G., Paterson,A.H., Ivarie,R., Brandstrom,M., Axelsson,E., Backstrom,N., Berlin,S., Webster,M.T., Pourquie,O., Reymond,A., Ucla,C., Antonarakis,S.E., Long,M., Emerson,J.J., Betran,E., Dupanloup,I., Kaessmann,H., Hinrichs,A.S., Bejerano,G., Furey,T.S., Harte,R.A., Raney,B., Siepel,A., Kent,W.J., Haussler,D., Eyras,E., Castelo,R., Abril,J.F., Castellano,S., Camara,F., Parra,G., Guigo,R., Bourque,G., Tesler,G., Pevzner,P.A., Smit,A., Fulton,L.A., Mardis,E.R., and Wilson,R.K. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695-716.

- Hinds,D.A., Stuve,L.L., Nilsen,G.B., Halperin,E., Eskin,E., Ballinger,D.G., Frazer,K.A., and Cox,D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072-1079.
- Hinshaw,J.E. (2000). Dynamin and its role in membrane fission. Annu. Rev. Cell Dev. Biol. *16*, 483-519.
- Hirschhorn,J.N. and Daly,M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95-108.
- Hodgkinson,C.A., Goldman,D., Jaeger,J., Persaud,S., Kane,J.M., Lipsky,R.H., and Malhotra,A.K. (2004). Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder. Am. J. Hum. Genet. *75*, 862-872.
- Hoeffer,A. and Pollin,W. (1970). Schizophrenia in the NAS-NRC panel of 15,909 veteran twin pairs. Arch. Gen. Psychiatry *23*, 469-477.
- Hoefgen,B., Schulze,T.G., Ohlraun,S., von Widdern,O., Hofels,S., Gross,M., Heidmann,V., Kovalenko,S., Eckermann,A., Kolsch,H., Metten,M., Zobel,A., Becker,T., Nothen,M.M., Propping,P., Heun,R., Maier,W., and Rietschel,M. (2005). The power of sample size and homogenous sampling: association between the 5-HTTLPR serotonin transporter polymorphism and major depressive disorder. Biol. Psychiatry *57*, 247-251.
- Hokfelt,T., Rehfeld,J.F., Skirboll,L., Ivemark,B., Goldstein,M., and Markey,K. (1980). Evidence for coexistence of dopamine and CCK in meso-limbic neurones. Nature *285*, 476-478.
- Holland,T. and Gosden,C. (1990). A balanced chromosomal translocation partially co-segregating with psychotic illness in a family. Psychiatry Res. *32*, 1-8.
- Holmes,J., Payton,A., Barrett,J., Harrington,R., McGuffin,P., Owen,M., Ollier,W., Worthington,J., Gill,M., Kirley,A., Hawi,Z., Fitzgerald,M., Asherson,P., Curran,S., Mill,J., Gould,A., Taylor,E., Kent,L., Craddock,N., and Thapar,A. (2002). Association of DRD4 in children with ADHD and comorbid conduct problems. Am. J. Med. Genet. *114*, 150-153.
- Holzman,P.S. (2000). Eye movements and the search for the essence of schizophrenia. Brain Res. Brain Res. Rev. *31*, 350-356.
- Horikawa,Y., Oda,N., Cox,N.J., Li,X., Orho-Melander,M., Hara,M., Hinokio,Y., Lindner,T.H., Mashima,H., Schwarz,P.E., Bosque-Plata,L., Horikawa,Y., Oda,Y., Yoshiuchi,I., Colilla,S., Polonsky,K.S., Wei,S., Concannon,P., Iwasaki,N., Schulze,J., Baier,L.J., Bogardus,C., Groop,L., Boerwinkle,E., Hanis,C.L., and Bell,G.I. (2000a). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat. Genet. *26*, 163-175.
- Horikawa,Y., Oda,N., Cox,N.J., Li,X., Orho-Melander,M., Hara,M., Hinokio,Y., Lindner,T.H., Mashima,H., Schwarz,P.E., Bosque-Plata,L., Horikawa,Y., Oda,Y., Yoshiuchi,I., Colilla,S., Polonsky,K.S., Wei,S., Concannon,P., Iwasaki,N., Schulze,J., Baier,L.J., Bogardus,C., Groop,L., Boerwinkle,E., Hanis,C.L., and Bell,G.I. (2000b). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat. Genet. *26*, 163-175.
- Hovatta,I., Varilo,T., Suvisaari,J., Terwilliger,J.D., Ollikainen,V., Arajarvi,R., Juvonen,H., Kokko-Sahin,M.L., Vaisanen,L., Mannila,H., Lonnqvist,J., and Peltonen,L. (1999). A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. Am. J. Hum. Genet. *65*, 1114-1124.
- Hranilovic,D., Stefulj,J., Schwab,S., Borrmann-Hassenbach,M., Albus,M., Jernej,B., and Wildenauer,D. (2004). Serotonin transporter promoter and intron 2 polymorphisms: relationship between allelic variants and gene expression. Biol. Psychiatry *55*, 1090-1094.
- Hsia,C.C. and McGinnis,W. (2003). Evolution of transcription factor function. Curr. Opin. Genet. Dev. *13*, 199-206.
- Huang,Y.Y., Cate,S.P., Battistuzzi,C., Oquendo,M.A., Brent,D., and Mann,J.J. (2004). An association between a functional polymorphism in the monoamine oxidase a gene promoter, impulsive traits and early abuse experiences. Neuropsychopharmacology *29*, 1498-1505.

- Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Huminiecki,L., Kasprzyk,A., Lehvaslaiho,H., Lijnzaad,P., Melsopp,C., Mongin,E., Pettett,R., Pocock,M., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., and Clamp,M. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38-41.
- Hughes,J.D., Estep,P.W., Tavazoie,S., and Church,G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. *296*, 1205-1214.
- Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M., Binder,V., Finkel,Y., Cortot,A., Modigliani,R., Laurent-Puig,P., Gower-Rousseau,C., Macry,J., Colombel,J.F., Sahbatou,M., and Thomas,G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599-603.
- Husi,H., Ward,M.A., Choudhary,J.S., Blackstock,W.P., and Grant,S.G. (2000). Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. Nat. Neurosci. *3*, 661-669.
- Ilusted,J., Scutt,L.E., and Bassett,A.S. (1998). Paternal transmission and anticipation in schizophrenia. Am. J. Med. Genet. *81*, 156-162.
- Huttunen,M.O. and Niskanen,P. (1978). Prenatal loss of father and psychiatric disorders. Arch. Gen. Psychiatry *35*, 429-431.
- Hwu,H.G., Liu,C.M., Fann,C.S., Ou-Yang,W.C., and Lee,S.F. (2003). Linkage of schizophrenia with chromosome 1q loci in Taiwanese families. Mol. Psychiatry *8*, 445-452.
- Igarashi,T. and Kaminuma,T. (1997). Development of a cell signaling networks database. Pac. Symp. Biocomput. 187-197.
- Ikeda,M., Iwata,N., Suzuki,T., Kitajima,T., Yamanouchi,Y., Kinoshita,Y., Inada,T., and Ozaki,N. (2004). Association of AKT1 with schizophrenia confirmed in a Japanese population. Biol. Psychiatry *56*, 698-700.
- Inada,T., Arinami,T., and Yagi,G. (1999). Association between a polymorphism in the promoter region of the dopamine D2 receptor gene and schizophrenia in Japanese subjects: replication and evaluation for antipsychotic-related features. Int. J. Neuropsychopharmcol. *2*, 181-186.
- Inoue,A. and Okabe,S. (2003). The dynamic organization of postsynaptic proteins: translocating molecules regulate synaptic function. Curr. Opin. Neurobiol. *13*, 332-340.
- Inoue,H., Tanizawa,Y., Wasson,J., Behn,P., Kalidas,K., Bernal-Mizrachi,E., Mueckler,M., Marshall,H., Donis-Keller,H., Crock,P., Rogers,D., Mikuni,M., Kumashiro,H., Higashi,K., Sobue,G., Oka,Y., and Permutt,M.A. (1998). A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). Nat. Genet. *20*, 143-148.
- Inoue,K. and Lupski,J.R. (2003). Genetics and genomics of behavioral and psychiatric disorders. Curr. Opin. Genet. Dev. *13*, 303-309.
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
- Ishiguro,H., Arinami,T., Saito,T., Akazawa,S., Enomoto,M., Mitushio,H., Fujishiro,H., Tada,K., Akimoto,Y., Mifune,H., Shiozuka,S., Hamaguchi,H., Toru,M., and Shibuya,H. (1998). Systematic search for variations in the tyrosine hydroxylase gene and their associations with schizophrenia, affective disorders, and alcoholism. Am. J. Med. Genet. *81*, 388-396.
- Ishiguro,H., Saito,T., Shibuya,H., Toru,M., and Arinami,T. (1999). No association between C-45T polymorphism in the Sp1 binding site of the promoter region of the cholecystokinin gene and alcoholism. Psychiatry Res. *85*, 209-213.
- Itokawa,M., Yamada,K., Iwayama-Shigeno,Y., Ishitsuka,Y., Detera-Wadleigh,S., and Yoshikawa,T. (2003a). Genetic analysis of a functional GRIN2A promoter (GT)n repeat in bipolar disorder pedigrees in humans. Neurosci. Lett. *345*, 53-56.

- Itokawa,M., Yamada,K., Yoshitsugu,K., Toyota,T., Suga,T., Ohba,H., Watanabe,A., Hattori,E., Shimizu,H., Kumakura,T., Ebihara,M., Meerabux,J.M., Toru,M., and Yoshikawa,T. (2003b). A microsatellite repeat in the promoter of the N-methyl-D-aspartate receptor 2A subunit (GRIN2A) gene suppresses transcriptional activity and correlates with chronic outcome in schizophrenia. Pharmacogenetics *13*, 271-278.
- James,H.M., Coller,J.K., Gillis,D., Bahnisch,J., Sallustio,B.C., and Somogyi,A.A. (2004a). A new simple diagnostic assay for the identification of the major CYP2D6 genotypes by DNA sequencing analysis. Int. J. Clin. Pharmacol. Ther. *42*, 719-723.
- James,R., Adams,R.R., Christie,S., Buchanan,S.R., Porteous,D.J., and Millar,J.K. (2004b). Disrupted in Schizophrenia 1 (DISC1) is a multicompartmentalized protein that predominantly localizes to mitochondria. Mol. Cell Neurosci. *26*, 112-122.
- Jareborg,N., Birney,E., and Durbin,R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. *9*, 815-824.
- Jareborg,N. and Durbin,R. (2000). Alfresco--a workbench for comparative genomic sequence analysis. Genome Res. *10*, 1148-1157.
- Jeffreys,A.J., Kauppi,L., and Neumann,R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. *29*, 217-222.
- Johnson,A.D., Wang,D., and Sadee,W. (2005). Polymorphisms affecting gene regulation and mRNA processing: broad implications for pharmacogenetics. Pharmacol. Ther. *106*, 19-38.
- Johnson,G.C., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A., Dudbridge,F., Twells,R.C., Payne,F., Hughes,W., Nutland,S., Stevens,H., Carr,P., Tuomilehto-Wolf,E., Tuomilehto,J., Gough,S.C., Clayton,D.G., and Todd,J.A. (2001). Haplotype tagging for the identification of common disease genes. Nat. Genet. *29*, 233-237.
- Johnson,M.P., Frescas,S.P., Oberlender,R., and Nichols,D.E. (1991). Synthesis and pharmacological examination of 1-(3-methoxy-4-methylphenyl)-2-aminopropane and 5-methoxy-6-methyl-2-aminoindan: similarities to 3,4-(methylenedioxy)methamphetamine (MDMA). J. Med. Chem. *34*, 1662-1668.
- Jonsson,E.G., Geijer,T., Gyllander,A., Terenius,L., and Sedvall,G.C. (1998). Failure to replicate an association between a rare allele of a tyrosine hydroxylase gene microsatellite and schizophrenia. Eur. Arch. Psychiatry Clin. Neurosci. *248*, 61-63.
- Jonsson,E.G., Nothen,M.M., Neidt,H., Forslund,K., Rylander,G., Mattila-Evenden,M., Asberg,M., Propping,P., and Sedvall,G.C. (1999). Association between a promoter polymorphism in the dopamine D2 receptor gene and schizophrenia. Schizophr. Res. *40*, 31-36.
- Joober,R., Benkelfat,C., Brisebois,K., Toulouse,A., Lafreniere,R.G., Turecki,G., Lal,S., Bloom,D., Labelle,A., Lalonde,P., Fortin,D., Alda,M., Palmour,R., and Rouleau,G.A. (1999). Lack of association between the hSKCa3 channel gene CAG polymorphism and schizophrenia. Am. J. Med. Genet. *88*, 154-157.
- Jorde,L.B. (2000). Linkage disequilibrium and the search for complex disease genes. Genome Res. *10*, 1435-1444.
- Jowett,B. (1898). Dialogues of Plato. (New York: D. Appleton and company).
- Joyce,J.N., Shane,A., Lexow,N., Winokur,A., Casanova,M.F., and Kleinman,J.E. (1993). Serotonin uptake sites and serotonin receptors are altered in the limbic system of schizophrenics. Neuropsychopharmacology *8*, 315-336.
- Kadonaga,J.T. (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell *116*, 247-257.
- Kaiser,R., Tremblay,P.B., Schmider,J., Henneken,M., Dettling,M., Muller-Oerlinghausen,B., Uebelhack,R., Roots,I., and Brockmoller,J. (2001). Serotonin transporter polymorphisms: no association with response to antipsychotic treatment, but associations with the schizoparanoid and residual subtypes of schizophrenia. Mol. Psychiatry *6*, 179-185.
- Kan,P.X., Popendikyte,V., Kaminsky,Z.A., Yolken,R.H., and Petronis,A. (2004). Epigenetic studies of genomic retroelements in major psychosis. Schizophr. Res. *67*, 95-106.

- Kane,J.M. and Freeman,H.L. (1994). Towards more effective antipsychotic treatment. Br. J. Psychiatry Suppl 22-31.
- Karayiorgou,M. and Gogos,J.A. (2004). The molecular genetics of the 22q11-associated schizophrenia. Brain Res. Mol. Brain Res. *132*, 95-104.
- Karayiorgou,M., Morris,M.A., Morrow,B., Shprintzen,R.J., Goldberg,R., Borrow,J., Gos,A., Nestadt,G., Wolyniec,P.S., and Lasseter,V.K. (1995). Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. Proc. Natl. Acad. Sci. U. S. A *92*, 7612-7616.
- Karlsson,H., Bachmann,S., Schroder,J., McArthur,J., Torrey,E.F., and Yolken,R.H. (2001). Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc. Natl. Acad. Sci. U. S. A. *98*, 4634-4639.
- Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D., and Kent,W.J. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51-54.
- Kato,T. (2001a). DNA polymorphisms and bipolar disorder. Am. J. Psychiatry *158*, 1169-1170.
- Kato,T. (2001b). Molecular genetics of bipolar disorder. Neurosci. Res. *40*, 105-113.
- Kato,T. and Kato,N. (2000). Mitochondrial dysfunction in bipolar disorder. Bipolar. Disord. *2*, 180-190.
- Kato,T., Kunugi,H., Nanko,S., and Kato,N. (2001). Mitochondrial DNA polymorphisms in bipolar disorder. J. Affect. Disord. *62*, 151-164.
- Kauppi,L., Sajantila,A., and Jeffreys,A.J. (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Hum. Mol. Genet. *12*, 33-40.
- Kawada,Y., Hattori,M., Dai,X.Y., and Nanko,S. (1995). Possible association between monoamine oxidase A gene and bipolar affective disorder. Am. J. Hum. Genet. *56*, 335-336.
- Kazazian,H.H.Jr., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G., and Antonarakis,S.E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164-166.
- Keikhaee,M.R., Fadai,F., Sargolzaee,M.R., Javanbakht,A., Najmabadi,H., and Ohadi,M. (2005). Association analysis of the dopamine transporter (DAT1)-67A/T polymorphism in bipolar disorder. Am. J. Med. Genet. B Neuropsychiatr. Genet. *135*, 47-49.
- Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V., and Wingender,E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. *31*, 3576-3579.
- Keller,C.A., Yuan,X., Panzanelli,P., Martin,M.L., Alldred,M., Sassoe-Pognetto,M., and Luscher,B. (2004). The gamma2 subunit of GABA(A) receptors is a substrate for palmitoylation by GODZ. J. Neurosci. *24*, 5881-5891.
- Kelley,S. (2000). Getting started with Acedb. Brief. Bioinform. *1*, 131-137.
- Kelsoe,J.R., Spence,M.A., Loetscher,E., Foguet,M., Sadovnick,A.D., Remick,R.A., Flodman,P., Khristich,J., Mroczkowski-Parker,Z., Brown,J.L., Masser,D., Ungerleider,S., Rapaport,M.H., Wishart,W.L., and Luebbert,H. (2001). A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. Proc. Natl. Acad. Sci. U. S. A. *98*, 585-590.
- Kennan,A., Aherne,A., and Humphries,P. (2005). Light in retinitis pigmentosa. Trends Genet. *21*, 103-110.
- Kenneson,A., Zhang,F., Hagedorn,C.H., and Warren,S.T. (2001). Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. Hum. Mol. Genet. *10*, 1449-1454.
- Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M., and Haussler,D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996-1006.

- Kerem,B., Rommens,J.M., Buchanan,J.A., Markiewicz,D., Cox,T.K., Chakravarti,A., Buchwald,M., and Tsui,L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073-1080.
- Kim,D.K., Lim,S.W., Lee,S., Sohn,S.E., Kim,S., Hahn,C.G., and Carroll,B.J. (2000). Serotonin transporter gene polymorphism and antidepressant response. Neuroreport *11*, 215-219.
- Kim,H.S., Park,J.S., Hong,S.J., Woo,M.S., Kim,S.Y., and Kim,K.S. (2003). Regulation of the tyrosine hydroxylase gene promoter by histone deacetylase inhibitors. Biochem. Biophys. Res. Commun. *312*, 950-957.
- Kim,S.J., Cox,N., Courchesne,R., Lord,C., Corsello,C., Akshoomoff,N., Guter,S., Leventhal,B.L., Courchesne,E., and Cook,E.H., Jr. (2002). Transmission disequilibrium mapping at the serotonin transporter gene (SLC6A4) region in autistic disorder. Mol. Psychiatry *7*, 278-288.
- King,M.C. and Wilson,A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.
- Kirchheiner,J., Brosen,K., Dahl,M.L., Gram,L.F., Kasper,S., Roots,I., Sjoqvist,F., Spina,E., and Brockmoller,J. (2001). CYP2D6 and CYP2C19 genotype-based dose recommendations for antidepressants: a first step towards subpopulation-specific dosages. Acta Psychiatr. Scand. *104*, 173-192.
- Kirkness,E.F., Bafna,V., Halpern,A.L., Levy,S., Remington,K., Rusch,D.B., Delcher,A.L., Pop,M., Wang,W., Fraser,C.M., and Venter,J.C. (2003). The dog genome: survey sequencing and comparative analysis. Science *301*, 1898-1903.
- Kirkwood,J. (2000). Sybase Adaptive Server Enterprise. (Welshpool: Kirkwood Associates).
- Kirov,G., Jones,I., McCandless,F., Craddock,N., and Owen,M.J. (1999). Family-based association studies of bipolar disorder with candidate genes involved in dopamine neurotransmission: DBH, DAT1, COMT, DRD2, DRD3 and DRD5. Mol. Psychiatry *4*, 558-565.
- Kitai,S.T., Shepard,P.D., Callaway,J.C., and Scroggs,R. (1999). Afferent modulation of dopamine neuron firing patterns. Curr. Opin. Neurobiol. *9*, 690-697.
- Klein,D.A. and Walsh,B.T. (2004). Eating disorders: clinical features and pathophysiology. Physiol Behav. *81*, 359-374.
- Klein,P.S. and Melton,D.A. (1996). A molecular mechanism for the effect of lithium on development. Proc. Natl. Acad. Sci. U. S. A *93*, 8455-8459.
- Knight,J.C. (2004). Allele-specific gene expression uncovered. Trends Genet. *20*, 113-116.
- Knight,J.C. (2005). Regulatory polymorphisms underlying complex disease traits. J. Mol. Med. *83*, 97-109.
- Knight,J.C., Keating,B.J., Rockett,K.A., and Kwiatkowski,D.P. (2003). In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat. Genet. *33*, 469-475.
- Knuppel,R., Dietze,P., Lehnberg,W., Frech,K., and Wingender,E. (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol. *1*, 191-198.
- Kobayashi,K., Morita,S., Sawada,H., Mizuguchi,T., Yamada,K., Nagatsu,I., Hata,T., Watanabe,Y., Fujita,K., and Nagatsu,T. (1995). Targeted disruption of the tyrosine hydroxylase locus results in severe catecholamine depletion and perinatal lethality in mice. J. Biol. Chem. *270*, 27235-27243.
- Kobayashi,S., Takashima,A., and Anzai,K. (1998). The dendritic translocation of translin protein in the form of BC1 RNA protein particles in developing rat hippocampal neurons in primary culture. Biochem. Biophys. Res. Commun. *253*, 448-453.
- Koenig,M., Hoffman,E.P., Bertelson,C.J., Monaco,A.P., Feener,C., and Kunkel,L.M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and. Cell *50*, 509-517.

- Koivisto,U.M., Palvimo,J.J., Janne,O.A., and Kontula,K. (1994). A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. Proc. Natl. Acad. Sci. U. S. A *91*, 10526-10530.
- Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R., and Chiaromonte,F. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. Genome Res. *14*, 700-707.
- Kopin,A.S., Lee,Y.M., McBride,E.W., Miller,L.J., Lu,M., Lin,H.Y., Kolakowski,L.F., Jr., and Beinborn,M. (1992). Expression cloning and characterization of the canine parietal cell gastrin receptor. Proc. Natl. Acad. Sci. U. S. A *89*, 3605-3609.
- Korf,I., Flicek,P., Duan,D., and Brent,M.R. (2001). Integrating genomic homology into gene structure prediction. Bioinformatics. *17 Suppl 1*, S140-S148.
- Koronyo-Hamaoui,M., Gak,E., Stein,D., Frisch,A., Danziger,Y., Leor,S., Michaelovsky,E., Laufer,N., Carel,C., Fennig,S., Mimouni,M., Apter,A., Goldman,B., Barkai,G., and Weizman,A. (2004). CAG repeat polymorphism within the KCNN3 gene is a significant contributor to susceptibility to anorexia nervosa: a case-control study of female patients and several ethnic groups in the Israeli Jewish population. Am. J. Med. Genet. B Neuropsychiatr. Genet. *131*, 76-80.
- Kraepelin,E. (1899). Psychiatrie: Ein Lehrbuch fur Studierende und Artze. (Leipzip: Barth Verlag).
- Krasnewich,D. and Gahl,W.A. (1997). Carbohydrate-deficient glycoprotein syndrome. Adv. Pediatr. *44*, 109-140.
- Kreahling,J. and Graveley,B.R. (2004). The origins and implications of Aluternative splicing. Trends Genet. *20*, 1-4.
- Kruglyak,L. and Nickerson,D.A. (2001). Variation is the spice of life. Nat. Genet. *27*, 234-236.
- Kunugi,H., Ishida,S., Kato,T., Tatsumi,M., Sakai,T., Hattori,M., Hirose,T., and Nanko,S. (1999). A functional polymorphism in the promoter region of monoamine oxidase-A gene and mood disorders. Mol. Psychiatry *4*, 393-395.
- Kunugi,H., Vallada,H.P., Sham,P.C., Hoda,F., Arranz,M.J., Li,T., Nanko,S., Murray,R.M., McGuffin,P., Owen,M., Gill,M., and Collier,D.A. (1997). Catechol-O-methyltransferase polymorphisms and schizophrenia: a transmission disequilibrium study in multiply affected families. Psychiatr. Genet. *7*, 97-101.
- Kustanovich,V., Ishii,J., Crawford,L., Yang,M., McGough,J.J., McCracken,J.T., Smalley,S.L., and Nelson,S.F. (2004). Transmission disequilibrium testing of dopamine-related candidate gene polymorphisms in ADHD: confirmation of association of ADHD with DRD4 and DRD5. Mol. Psychiatry *9*, 711-717.
- Labelle,C. and Leclerc,N. (2000). Exogenous BDNF, NT-3 and NT-4 differentially regulate neurite outgrowth in cultured hippocampal neurons. Brain Res. Dev. Brain Res. *123*, 1-11.
- Lachman,H.M., Morrow,B., Shprintzen,R., Veit,S., Parsia,S.S., Faedda,G., Goldberg,R., Kucherlapati,R., and Papolos,D.F. (1996a). Association of codon 108/158 catechol-O-methyltransferase gene polymorphism with the psychiatric manifestations of velo-cardio-facial syndrome. Am. J. Med. Genet. *67*, 468-472.
- Lachman,H.M., Papolos,D.F., Saito,T., Yu,Y.M., Szumlanski,C.L., and Weinshilboum,R.M. (1996b). Human catechol-O-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. Pharmacogenetics *6*, 243-250.
- Lai,C.S., Fisher,S.E., Hurst,J.A., Vargha-Khadem,F., and Monaco,A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519-523.
- Lai,C.S., Gerrelli,D., Monaco,A.P., Fisher,S.E., and Copp,A.J. (2003). FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. Brain *126*, 2455-2462.
- Lander,E. and Kruglyak,L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241-247.
- Lander,E.S. (1996). The new genomics: global views of biology. Science *274*, 536-539.

- Lander,E.S. and Schork,N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037-2048.
- Lange,K.J. and McInnis,M.G. (2002). Studies of anticipation in bipolar affective disorder. CNS. Spectr. *7*, 196-202.
- Larsson,L.I. and Rehfeld,J.F. (1979). Localization and molecular heterogeneity of cholecystokinin in the central and peripheral nervous system. Brain Res. *165*, 201-218.
- Lasky-Su,J.A., Faraone,S.V., Glatt,S.J., and Tsuang,M.T. (2005). Meta-analysis of the association between two polymorphisms in the serotonin transporter gene and affective disorders. Am. J. Med. Genet. B Neuropsychiatr. Genet. *133*, 110-115.
- Laurent,C., Niehaus,D., Bauche,S., Levinson,D.F., Soubigou,S., Pimstone,S., Hayden,M., Mbanga,I., Emsley,R., Deleuze,J.F., and Mallet,J. (2003). CAG repeat polymorphisms in KCNN3 (HSKCa3) and PPP2R2B show no association or linkage to schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *116*, 45-50.
- Laurent,C., Zander,C., Thibaut,F., Bonnet-Brilhault,F., Chavand,O., Jay,M., Samolyk,D., Petit,M., Martinez,M., Campion,D., Neri,C., Mallet,J., and Cann,H. (1998). Anticipation in schizophrenia: no evidence of expanded CAG/CTG repeat sequences in French families and sporadic cases. Am. J. Med. Genet. *81*, 342-346.
- Law,A.J., Shannon,W.C., Hyde,T.M., Kleinman,J.E., and Harrison,P.J. (2004). Neuregulin-1 (NRG-1) mRNA and protein in the adult human brain. Neuroscience *127*, 125-136.
- Laws,S.M., Hone,E., Taddei,K., Harper,C., Dean,B., McClean,C., Masters,C., Lautenschlager,N., Gandy,S.E., and Martins,R.N. (2002). Variation at the APOE -491 promoter locus is associated with altered brain levels of apolipoprotein E. Mol. Psychiatry *7*, 886-890.
- Le Hellard,S., Ballereau,S.J., Visscher,P.M., Torrance,H.S., Pinson,J., Morris,S.W., Thomson,M.L., Semple,C.A., Muir,W.J., Blackwood,D.H., Porteous,D.J., and Evans,K.L. (2002). SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. Nucleic Acids Res. *30*, e74.
- Le Hellard,S., Semple,C.A., Morris,S.W., Porteous,D.J., and Evans,K.L. (2001). Physical mapping: integrating computational and molecular genetic data. Ann. Hum. Genet. *65*, 221-228.
- Lebeda,F.J. (2004). BotDB: A database resource for the clostridial neurotoxins. Mov Disord. *19 Suppl 8*, S35-S41.
- Lee,F.J., Xue,S., Pei,L., Vukusic,B., Chery,N., Wang,Y., Wang,Y.T., Niznik,H.B., Yu,X.M., and Liu,F. (2002). Dual regulation of NMDA receptor functions by direct protein-protein interactions with the dopamine D1 receptor. Cell *111*, 219-230.
- Lemonde,S., Turecki,G., Bakish,D., Du,L., Hrdina,P.D., Bown,C.D., Sequeira,A., Kushwaha,N., Morris,S.J., Basak,A., Ou,X.M., and Albert,P.R. (2003). Impaired repression at a 5-hydroxytryptamine 1A receptor gene polymorphism associated with major depression and suicide. J. Neurosci. *23*, 8788-8799.
- Leonard,S., Gault,J., Hopkins,J., Logel,J., Vianzon,R., Short,M., Drebing,C., Berger,R., Venn,D., Sirota,P., Zerbe,G., Olincy,A., Ross,R.G., Adler,L.E., and Freedman,R. (2002). Association of promoter variants in the alpha7 nicotinic acetylcholine receptor subunit gene with an inhibitory deficit found in schizophrenia. Arch. Gen. Psychiatry *59*, 1085-1096.
- Lesch,K.P., Balling,U., Gross,J., Strauss,K., Wolozin,B.L., Murphy,D.L., and Riederer,P. (1994). Organization of the human serotonin transporter gene. J. Neural Transm. Gen. Sect. *95*, 157-162.
- Lesch,K.P., Bengel,D., Heils,A., Sabol,S.Z., Greenberg,B.D., Petri,S., Benjamin,J., Muller,C.R., Hamer,D.H., and Murphy,D.L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science *274*, 1527-1531.
- Lesch,K.P. and Mossner,R. (1998). Genetically driven variation in serotonin uptake: is there a link to affective spectrum, neurodevelopmental, and neurodegenerative disorders? Biol. Psychiatry *44*, 179-192.
- Lettice,L.A., Horikoshi,T., Heaney,S.J., van Baren,M.J., van der Linde,H.C., Breedveld,G.J., Joosse,M., Akarsu,N., Oostra,B.A., Endo,N., Shibata,M., Suzuki,M., Takahashi,E.,

Shinka,T., Nakahori,Y., Ayusawa,D., Nakabayashi,K., Scherer,S.W., Heutink,P., Hill,R.E., and Noji,S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc. Natl. Acad. Sci. U. S. A. *99*, 7548-7553.

- Leucht,S., Wahlbeck,K., Hamann,J., and Kissling,W. (2003). New generation antipsychotics versus low-potency conventional antipsychotics: a systematic review and meta-analysis. Lancet *361*, 1581-1589.
- Leveugle,M., Prat,K., Perrier,N., Birnbaum,D., and Coulier,F. (2003). ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic Acids Res. *31*, 63-67.
- Levine,M. and Tjian,R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.
- Levitan,R.D., Masellis,M., Basile,V.S., Lam,R.W., Kaplan,A.S., Davis,C., Muglia,P., Mackenzie,B., Tharmalingam,S., Kennedy,S.H., Macciardi,F., and Kennedy,J.L. (2004a). The dopamine-4 receptor gene associated with binge eating and weight gain in women with seasonal affective disorder: an evolutionary perspective. Biol. Psychiatry *56*, 665-669.
- Levitan,R.D., Masellis,M., Lam,R.W., Muglia,P., Basile,V.S., Jain,U., Kaplan,A.S., Tharmalingam,S., Kennedy,S.H., and Kennedy,J.L. (2004b). Childhood inattention and dysphoria and adult obesity associated with the dopamine D4 receptor gene in overeating women with seasonal affective disorder. Neuropsychopharmacology *29*, 179-186.
- Levy,S. and Hannenhalli,S. (2002). Identification of transcription factor binding sites in the human genome sequence. Mamm. Genome *13*, 510-514.
- Lewis,C.M., Levinson,D.F., Wise,L.H., DeLisi,L.E., Straub,R.E., Hovatta,I., Williams,N.M., Schwab,S.G., Pulver,A.E., Faraone,S.V., Brzustowicz,L.M., Kaufmann,C.A., Garver,D.L., Gurling,H.M., Lindholm,E., Coon,H., Moises,H.W., Byerley,W., Shaw,S.H., Mesen,A., Sherrington,R., O'Neill,F.A., Walsh,D., Kendler,K.S., Ekelund,J., Paunio,T., Lonnqvist,J., Peltonen,L., O'Donovan,M.C., Owen,M.J., Wildenauer,D.B., Maier,W., Nestadt,G., Blouin,J.L., Antonarakis,S.E., Mowry,B.J., Silverman,J.M., Crowe,R.R., Cloninger,C.R., Tsuang,M.T., Malaspina,D., Harkavy-Friedman,J.M., Svrakic,D.M., Bassett,A.S., Holcomb,J., Kalsi,G., McQuillin,A., Brynjolfson,J., Sigmundsson,T., Petursson,H., Jazin,E., Zoega,T., and Helgason,T. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. Am. J. Hum. Genet. *73*, 34-48.
- Li,T., Arranz,M., Aitchison,K.J., Bryant,C., Liu,X., Kerwin,R.W., Murray,R., Sham,P., and Collier,D.A. (1998a). Case-control, haplotype relative risk and transmission disequilibrium analysis of a dopamine D2 receptor functional promoter polymorphism in schizophrenia. Schizophr. Res. *32*, 87-92.
- Li,T., Holmes,C., Sham,P.C., Vallada,H., Birkett,J., Kirov,G., Lesch,K.P., Powell,J., Lovestone,S., and Collier,D. (1997a). Allelic functional variation of serotonin transporter expression is a susceptibility factor for late onset Alzheimer's disease. Neuroreport *8*, 683-686.
- Li,T., Hu,X., Chandy,K.G., Fantino,E., Kalman,K., Gutman,G., Gargus,J.J., Freeman,B., Murray,R.M., Dawson,E., Liu,X., Bruinvels,A.T., Sham,P.C., and Collier,D.A. (1998b). Transmission disequilibrium analysis of a triplet repeat within the hKCa3 gene using family trios with schizophrenia. Biochem. Biophys. Res. Commun. *251*, 662-665.
- Li,T., Liu,X., Sham,P.C., Aitchison,K.J., Cai,G., Arranz,M.J., Deng,H., Liu,J., Kirov,G., Murray,R.M., and Collier,D.A. (1999). Association analysis between dopamine receptor genes and bipolar affective disorder. Psychiatry Res. *86*, 193-201.
- Li,T., Ma,X., Sham,P.C., Sun,X., Hu,X., Wang,Q., Meng,H., Deng,W., Liu,X., Murray,R.M., and Collier,D.A. (2004a). Evidence for association between novel polymorphisms in the PRODH gene and schizophrenia in a Chinese population. Am. J. Med. Genet. B Neuropsychiatr. Genet. *129*, 13-15.
- Li,T., Sham,P.C., Vallada,H., Xie,T., Tang,X., Murray,R.M., Liu,X., and Collier,D.A. (1996). Preferential transmission of the high activity allele of COMT in schizophrenia. Psychiatr. Genet. *6*, 131-133.
- Li,T., Stefansson,H., Gudfinnsson,E., Cai,G., Liu,X., Murray,R.M., Steinthorsdottir,V., Januel,D., Gudnadottir,V.G., Petursson,H., Ingason,A., Gulcher,J.R., Stefansson,K., and

Collier,D.A. (2004b). Identification of a novel neuregulin 1 at-risk haplotype in Han schizophrenia Chinese patients, but no association with the Icelandic/Scottish risk haplotype. Mol. Psychiatry *9*, 698-704.

- Li,T., Vallada,H.P., Liu,X., Xie,T., Tang,X., Zhao,J., O'Donovan,M.C., Murray,R.M., Sham,P.C., and Collier,D.A. (1998c). Analysis of CAG/CTG repeat size in Chinese subjects with schizophrenia and bipolar affective disorder using the repeat expansion detection method. Biol. Psychiatry *44*, 1160-1165.
- Li,T., Vallada,H.P., Liu,X., Xie,T., Tang,X., Zhao,J., O'Donovan,M.C., Murray,R.M., Sham,P.C., and Collier,D.A. (1998d). Analysis of CAG/CTG repeat size in Chinese subjects with schizophrenia and bipolar affective disorder using the repeat expansion detection method. Biol. Psychiatry *44*, 1160-1165.
- Li,T., Xu,K., Deng,H., Cai,G., Liu,J., Liu,X., Wang,R., Xiang,X., Zhao,J., Murray,R.M., Sham,P.C., and Collier,D.A. (1997b). Association analysis of the dopamine D4 gene exon III VNTR and heroin abuse in Chinese subjects. Mol. Psychiatry *2*, 413-416.
- Li,Y.J., Fu,X.H., Liu,D.P., and Liang,C.C. (2004c). Opening the chromatin for transcription. Int. J. Biochem. Cell Biol. *36*, 1411-1423.
- Lim,L.C., Nothen,M.M., Korner,J., Rietschel,M., Castle,D., Hunt,N., Propping,P., Murray,R., and Gill,M. (1994). No evidence of association between dopamine D4 receptor variants and bipolar affective disorder. Am. J. Med. Genet. *54*, 259-263.
- Lim,L.C., Powell,J., Sham,P., Castle,D., Hunt,N., Murray,R., and Gill,M. (1995). Evidence for a genetic association between alleles of monoamine oxidase A gene and bipolar affective disorder. Am. J. Med. Genet. *60*, 325-331.
- Lin,S., Jiang,S., Wu,X., Qian,Y., Wang,D., Tang,G., and Gu,N. (2000). Association analysis between mood disorder and monoamine oxidase gene. Am. J. Med. Genet. *96*, 12-14.
- Lindblad,K., Nylander,P.O., Zander,C., Yuan,Q.P., Stahle,L., Engstrom,C., Balciuniene,J., Pettersson,U., Breschel,T., McInnis,M., Ross,C.A., Adolfsson,R., and Schalling,M. (1998). Two commonly expanded CAG/CTG repeat loci: involvement in affective disorders? Mol. Psychiatry *3*, 405-410.
- Liu,F., Wan,Q., Pristupa,Z.B., Yu,X.M., Wang,Y.T., and Niznik,H.B. (2000). Direct protein-protein coupling enables cross-talk between dopamine D5 and gamma-aminobutyric acid A receptors. Nature *403*, 274-280.
- Liu,H., Abecasis,G.R., Heath,S.C., Knowles,A., Demars,S., Chen,Y.J., Roos,J.L., Rapoport,J.L., Gogos,J.A., and Karayiorgou,M. (2002a). Genetic variation in the 22q11 locus and susceptibility to schizophrenia. Proc. Natl. Acad. Sci. U. S. A. *99*, 16859-16864.
- Liu,H., Heath,S.C., Sobin,C., Roos,J.L., Galke,B.L., Blundell,M.L., Lenane,M., Robertson,B., Wijsman,E.M., Rapoport,J.L., Gogos,J.A., and Karayiorgou,M. (2002b). Genetic variation at the 22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. Proc. Natl. Acad. Sci. U. S. A. *99*, 3717-3722.
- Liu,W., Gu,N., Feng,G., Li,S., Bai,S., Zhang,J., Shen,T., Xue,H., Breen,G., St Clair,D., and He,L. (1999). Tentative association of the serotonin transporter with schizophrenia and unipolar depression but not with bipolar disorder in Han Chinese. Pharmacogenetics *9*, 491-495.
- Lo,H.S., Wang,Z., Hu,Y., Yang,H.H., Gere,S., Buetow,K.H., and Lee,M.P. (2003). Allelic variation in gene expression is common in the human genome. Genome Res. *13*, 1855-1862.
- Lohmueller,K.E., Pearce,C.L., Pike,M., Lander,E.S., and Hirschhorn,J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. *33*, 177-182.
- Lohr,U., Yussa,M., and Pick,L. (2001). Drosophila fushi tarazu. a gene on the border of homeotic function. Curr. Biol. *11*, 1403-1412.
- Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M., and Frazer,K.A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136-140.

- Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I., and Rubin,E.M. (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. *12*, 832-839.
- Lotta,T., Vidgren,J., Tilgmann,C., Ulmanen,I., Melen,K., Julkunen,I., and Taskinen,J. (1995). Kinetics of human soluble and membrane-bound catechol O-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme. Biochemistry *34*, 4202-4210.
- Lovejoy,E.A., Scott,A.C., Fiskerstrand,C.E., Bubb,V.J., and Quinn,J.P. (2003). The serotonin transporter intronic VNTR enhancer correlated with a predisposition to affective disorders has distinct regulatory elements within the domain based on the primary DNA sequence of the repeat unit. Eur. J. Neurosci. *17*, 417-420.
- Lowe,N., Kirley,A., Hawi,Z., Sham,P., Wickham,H., Kratochvil,C.J., Smith,S.D., Lee,S.Y., Levy,F., Kent,L., Middle,F., Rohde,L.A., Roman,T., Tahir,E., Yazgan,Y., Asherson,P., Mill,J., Thapar,A., Payton,A., Todd,R.D., Stephens,T., Ebstein,R.P., Manor,I., Barr,C.L., Wigg,K.G., Sinke,R.J., Buitelaar,J.K., Smalley,S.L., Nelson,S.F., Biederman,J., Faraone,S.V., and Gill,M. (2004). Joint analysis of the DRD5 marker concludes association with attention-deficit/hyperactivity disorder confined to the predominantly inattentive and combined subtypes. Am. J. Hum. Genet. *74*, 348-356.
- Ludwig,M.Z. (2002). Functional evolution of noncoding DNA. Curr. Opin. Genet. Dev. *12*, 634-639.
- Luo,X., Klempan,T.A., Lappalainen,J., Rosenheck,R.A., Charney,D.S., Erdos,J., van Kammen,D.P., Kranzler,H.R., Kennedy,J.L., and Gelernter,J. (2004). NOTCH4 gene haplotype is associated with schizophrenia in African Americans. Biol. Psychiatry *55*, 112-117.
- Ma,L., Liu,Y., Ky,B., Shughrue,P.J., Austin,C.P., and Morris,J.A. (2002). Cloning and characterization of Disc1, the mouse ortholog of DISC1 (Disrupted-in-Schizophrenia 1). Genomics *80*, 662-672.
- Machon,O., van den Bout,C.J., Backman,M., Rosok,O., Caubit,X., Fromm,S.H., Geronimo,B., and Krauss,S. (2002). Forebrain-specific promoter/enhancer D6 derived from the mouse Dach1 gene controls expression in neural stem cells. Neuroscience *112*, 951-966.
- Maher,B.S., Marazita,M.L., Ferrell,R.E., and Vanyukov,M.M. (2002). Dopamine system genes and attention deficit hyperactivity disorder: a meta-analysis. Psychiatr. Genet. *12*, 207-215.
- Malhotra,A.K., Goldman,D., Mazzanti,C., Clifton,A., Breier,A., and Pickar,D. (1998). A functional serotonin transporter (5-HTT) polymorphism is associated with psychosis in neuroleptic-free schizophrenics. Mol. Psychiatry *3*, 328-332.
- Malhotra,A.K., Murphy,G.M., Jr., and Kennedy,J.L. (2004). Pharmacogenetics of psychotropic drug response. Am. J. Psychiatry *161*, 780-796.
- Manki,H., Kanba,S., Muramatsu,T., Higuchi,S., Suzuki,E., Matsushita,S., Ono,Y., Chiba,H., Shintani,F., Nakamura,M., Yagi,G., and Asai,M. (1996). Dopamine D2, D3 and D4 receptor and transporter gene polymorphisms and mood disorders. J. Affect. Disord. *40*, 7-13.
- Mannisto,P.T. and Kaakkola,S. (1999). Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors. Pharmacol. Rev. *51*, 593-628.
- Marchal-Victorion,S., Vionnet,N., Escrieut,C., Dematos,F., Dina,C., Dufresne,M., Vaysse,N., Pradayrol,L., Froguel,P., and Fourmy,D. (2002). Genetic, pharmacological and functional analysis of cholecystokinin-1 and cholecystokinin-2 receptor polymorphism in type 2 diabetes and obese patients. Pharmacogenetics *12*, 23-30.
- Margulies,E.H., Blanchette,M., Haussler,D., and Green,E.D. (2003). Identification and characterization of multi-species conserved sequences. Genome Res. *13*, 2507-2518.
- Martin,S.L., Blackmon,B.P., Rajagopalan,R., Houfek,T.D., Sceeles,R.G., Denn,S.O., Mitchell,T.K., Brown,D.E., Wing,R.A., and Dean,R.A. (2002). MagnaportheDB: a federated solution for integrating physical and genetic map data with BAC end derived sequences for the rice blast fungus Magnaporthe grisea. Nucleic Acids Res. *30*, 121-124.

- Massat,I., Souery,D., Del Favero,J., Van Gestel,S., Serretti,A., Macciardi,F., Smeraldi,E., Kaneva,R., Adolfsson,R., Nylander,P.O., Blackwood,D., Muir,W., Papadimitriou,G.N., Dikeos,D., Oruc,L., Segman,R.H., Ivezic,S., Aschauer,H., Ackenheil,M., Fuchshuber,S., Dam,H., Jakovljevic,M., Peltonen,L., Hilger,C., Hentges,F., Staner,L., Milanova,V., Jazin,E., Lerer,B., Van Broeckhoven,C., and Mendlewicz,J. (2002). Positive association of dopamine D2 receptor polymorphism with bipolar affective disorder in a European Multicenter Association Study of affective disorders. Am. J. Med. Genet. *114*, 177-185.
- Matsumoto,M., Weickert,C.S., Akil,M., Lipska,B.K., Hyde,T.M., Herman,M.M., Kleinman,J.E., and Weinberger,D.R. (2003a). Catechol O-methyltransferase mRNA expression in human and rat brain: evidence for a role in cortical neuronal function. Neuroscience *116*, 127-137.
- Matsumoto,M., Weickert,C.S., Beltaifa,S., Kolachana,B., Chen,J., Hyde,T.M., Herman,M.M., Weinberger,D.R., and Kleinman,J.E. (2003b). Catechol O-methyltransferase (COMT) mRNA expression in the dorsolateral prefrontal cortex of patients with schizophrenia. Neuropsychopharmacology *28*, 1521-1530.
- Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V., Kloos,D.U., Land,S., Lewicki-Potapov,B., Michael,H., Munch,R., Reuter,I., Rotert,S., Saxel,H., Scheer,M., Thiele,S., and Wingender,E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374-378.
- Max-Muller (1998). Sacred books of the East. (London: Routhledge).
- Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S., and Dubchak,I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. *16*, 1046-1047.
- McCracken,J.T., Smalley,S.L., McGough,J.J., Crawford,L., Del'Homme,M., Cantor,R.M., Liu,A., and Nelson,S.F. (2000). Evidence for linkage of a tandem duplication polymorphism upstream of the dopamine D4 receptor gene (DRD4) with attention deficit hyperactivity disorder (ADHD). Mol. Psychiatry *5*, 531-536.
- McDermott,D.H., Zimmerman,P.A., Guignard,F., Kleeberger,C.A., Leitman,S.F., and Murphy,P.M. (1998). CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). Lancet *352*, 866-870.
- McGinnis,R.E., Fox,H., Yates,P., Cameron,L.A., Barnes,M.R., Gray,I.C., Spurr,N.K., Hurko,O., and St Clair,D. (2001). Failure to confirm NOTCH4 association with schizophrenia in a large population-based sample from Scotland. Nat. Genet. *28*, 128-129.
- McGorry,P.D., Mihalopoulos,C., Henry,L., Dakis,J., Jackson,H.J., Flaum,M., Harrigan,S., McKenzie,D., Kulkarni,J., and Karoly,R. (1995). Spurious precision: procedural validity of diagnostic assessment in psychotic disorders. Am. J. Psychiatry *152*, 220-223.
- McInnes,L.A., Service,S.K., Reus,V.I., Barnes,G., Charlat,O., Jawahar,S., Lewitzky,S., Yang,Q., Duong,Q., Spesny,M., Araya,C., Araya,X., Gallegos,A., Meza,L., Molina,J., Ramirez,R., Mendez,R., Silva,S., Fournier,E., Batki,S.L., Mathews,C.A., Neylan,T., Glatt,C.E., Escamilla,M.A., Luo,D., Gajiwala,P., Song,T., Crook,S., Nguyen,J.B., Roche,E., Meyer,J.M., Leon,P., Sandkuijl,L.A., Freimer,N.B., and Chen,H. (2001). Fine-scale mapping of a locus for severe bipolar mood disorder on chromosome 18p11.3 in the Costa Rican population. Proc. Natl. Acad. Sci. U. S. A. *98*, 11485-11490.
- McInnis,M.G. (1996). Anticipation: an old idea in new genes. Am. J. Hum. Genet. *59*, 973-979.
- McInnis,M.G., Breschel,T.S., Margolis,R.L., Chellis,J., MacKinnon,D.F., McMahon,F.J., Simpson,S.G., Lan,T.H., Chen,H., Ross,C.A., and DePaulo,J.R. (1999a). Family-based association analysis of the hSKCa3 potassium channel gene in bipolar disorder. Mol. Psychiatry *4*, 217-219.
- McInnis,M.G., Dick,D.M., Willour,V.L., Avramopoulos,D., MacKinnon,D.F., Simpson,S.G., Potash,J.B., Edenberg,H.J., Bowman,E.S., McMahon,F.J., Smiley,C., Chellis,J.L., Huo,Y., Diggs,T., Meyer,E.T., Miller,M., Matteini,A.T., Rau,N.L., DePaulo,J.R., Gershon,E.S., Badner,J.A., Rice,J.P., Goate,A.M., Detera-Wadleigh,S.D., Nurnberger,J.I., Reich,T.,

Zandi,P.P., and Foroud,T.M. (2003). Genome-wide scan and conditional analysis in bipolar disorder: evidence for genomic interaction in the National Institute of Mental Health genetics initiative bipolar pedigrees. Biol. Psychiatry *54*, 1265-1273.

- McInnis,M.G., McMahon,F.J., Chase,G.A., Simpson,S.G., Ross,C.A., and DePaulo,J.R., Jr. (1993). Anticipation in bipolar affective disorder. Am. J. Hum. Genet. *53*, 385-390.
- McInnis,M.G., McMahon,F.J., Crow,T., Ross,C.A., and DeLisi,L.E. (1999b). Anticipation in schizophrenia: a review and reconsideration. Am. J. Med. Genet. *88*, 686-693.
- McIntyre,A., Summersgill,B., Jafer,O., Rodriguez,S., Zafarana,G., Oosterhuis,J.W., Gillis,A.J., Looijenga,L., Cooper,C., Huddart,R., Clark,J., and Shipley,J. (2004). Defining minimum genomic regions of imbalance involved in testicular germ cell tumors of adolescents and adults through genome wide microarray analysis of cDNA clones. Oncogene *23*, 9142-9147.
- McMahon,F.J., Chen,Y.S., Patel,S., Kokoszka,J., Brown,M.D., Torroni,A., DePaulo,J.R., and Wallace,D.C. (2000). Mitochondrial DNA sequence diversity in bipolar affective disorder. Am. J. Psychiatry *157*, 1058-1064.
- McMahon,F.J., Stine,O.C., Meyers,D.A., Simpson,S.G., and DePaulo,J.R. (1995). Patterns of maternal transmission in bipolar affective disorder. Am. J. Hum. Genet. *56*, 1277-1286.
- McQuillin,A., Lawrence,J., Curtis,D., Kalsi,G., Smyth,C., Hannesdottir,S., and Gurling,H. (1999). Adjacent genetic markers on chromosome 11p15.5 at or near the tyrosine hydroxylase locus that show population linkage disequilibrium with each other do not show allelic association with bipolar affective disorder. Psychol. Med. *29*, 1449-1454.
- Meloni,R., Laurent,C., Campion,D., Ben Hadjali,B., Thibaut,F., Dollfus,S., Petit,M., Samolyk,D., Martinez,M., and Poirier,M.F. (1995). A rare allele of a microsatellite located in the tyrosine hydroxylase gene found in schizophrenic patients. C. R. Acad. Sci. III *318*, 803-809.
- Mendlewicz,J., Lindbald,K., Souery,D., Mahieu,B., Nylander,P.O., De Bruyn,A., Zander,C., Engstrom,C., Adolfsson,R., Van Broeckhoven,C., Schalling,M., and Lipp,O. (1997). Expanded trinucleotide CAG repeats in families with bipolar affective disorder. Biol. Psychiatry *42*, 1115-1122.
- Merikangas,K.R. and Risch,N. (2003). Will the genomics revolution revolutionize psychiatry? Am. J. Psychiatry *160*, 625-635.
- Michelhaugh,S.K., Fiskerstrand,C., Lovejoy,E., Bannon,M.J., and Quinn,J.P. (2001). The dopamine transporter gene (SLC6A3) variable number of tandem repeats domain enhances transcription in dopamine neurons. J. Neurochem. *79*, 1033-1038.
- Miki,Y., Swensen,J., Shattuck-Eidens,D., Futreal,P.A., Harshman,K., Tavtigian,S., Liu,Q., Cochran,C., Bennett,L.M., and Ding,W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66-71.
- Mill,J., Asherson,P., Browes,C., D'Souza,U., and Craig,I. (2002). Expression of the dopamine transporter gene is regulated by the 3' UTR VNTR: Evidence from brain and lymphocytes using quantitative RT-PCR. Am. J. Med. Genet. *114*, 975-979.
- Mill,J., Curran,S., Richards,S., Taylor,E., and Asherson,P. (2004). Polymorphisms in the dopamine D5 receptor (DRD5) gene and ADHD. Am. J. Med. Genet. B Neuropsychiatr. Genet. *125*, 38-42.
- Mill,J., Fisher,N., Curran,S., Richards,S., Taylor,E., and Asherson,P. (2003). Polymorphisms in the dopamine D4 receptor gene and attention-deficit hyperactivity disorder. Neuroreport *14*, 1463-1466.
- Millar,J.K., Christie,S., and Porteous,D.J. (2003). Yeast two-hybrid screens implicate DISC1 in brain development and function. Biochem. Biophys. Res. Commun. *311*, 1019-1025.
- Millar,J.K., Christie,S., Semple,C.A., and Porteous,D.J. (2000a). Chromosomal location and genomic structure of the human translin-associated factor X gene (TRAX; TSNAX) revealed by intergenic splicing to DISC1, a gene disrupted by a translocation segregating with schizophrenia. Genomics *67*, 69-77.

- Millar,J.K., James,R., Brandon,N.J., and Thomson,P.A. (2004). DISC1 and DISC2: discovering and dissecting molecular mechanisms underlying psychiatric illness. Ann. Med. *36*, 367-378.
- Millar,J.K., Wilson-Annan,J.C., Anderson,S., Christie,S., Taylor,M.S., Semple,C.A., Devon,R.S., Clair,D.M., Muir,W.J., Blackwood,D.H., and Porteous,D.J. (2000b). Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum. Mol. Genet. *9*, 1415-1423.
- Miller,G., Fuchs,R., and Lai,E. (1997). IMAGE cDNA clones, UniGene clustering, and ACeDB: an integrated resource for expressed sequence information. Genome Res. *7*, 1027-1032.
- Miller,G.M. and Madras,B.K. (2002). Polymorphisms in the 3'-untranslated region of human and monkey dopamine transporter genes affect reporter gene expression. Mol. Psychiatry *7*, 44-55.
- Miller,L.J., Holicky,E.L., Ulrich,C.D., and Wieben,E.D. (1995). Abnormal processing of the human cholecystokinin receptor gene in association with gallstones and obesity. Gastroenterology *109*, 1375-1380.
- Miller,M.J., Rauer,H., Tomita,H., Rauer,H., Gargus,J.J., Gutman,G.A., Cahalan,M.D., and Chandy,K.G. (2001). Nuclear localization and dominant-negative suppression by a mutant SKCa3 N-terminal channel fragment identified in a patient with schizophrenia. J. Biol. Chem. *276*, 27753-27756.
- Mimmack,M.L., Ryan,M., Baba,H., Navarro-Ruiz,J., Iritani,S., Faull,R.L., McKenna,P.J., Jones,P.B., Arai,H., Starkey,M., Emson,P.C., and Bahn,S. (2002). Gene expression analysis in schizophrenia: reproducible up-regulation of several members of the apolipoprotein L family located in a high-susceptibility locus for schizophrenia on chromosome 22. Proc. Natl. Acad. Sci. U. S. A. *99*, 4680-4685.
- Minogue,S., Anderson,J.S., Waugh,M.G., dos,S.M., Corless,S., Cramer,R., and Hsuan,J.J. (2001). Cloning of a human type II phosphatidylinositol 4-kinase reveals a novel lipid kinase family. J. Biol. Chem. *276*, 16635-16640.
- Minov,C., Baghai,T.C., Schule,C., Zwanzger,P., Schwarz,M.J., Zill,P., Rupprecht,R., and Bondy,B. (2001). Serotonin-2A-receptor and -transporter polymorphisms: lack of association in patients with major depression. Neurosci. Lett. *303*, 119-122.
- Mirnics,K., Middleton,F.A., Stanwood,G.D., Lewis,D.A., and Levitt,P. (2001). Disease-specific changes in regulator of G-protein signaling 4 (RGS4) expression in schizophrenia. Mol. Psychiatry *6*, 293-301.
- Misbahuddin,A., Placzek,M.R., Chaudhuri,K.R., Wood,N.W., Bhatia,K.P., and Warner,T.T. (2002). A polymorphism in the dopamine receptor DRD5 is associated with blepharospasm. Neurology *58*, 124-126.
- Mishra,S. and Beaulieu,A. (2002). Mastering Oracle SQL. (Sebastopol, CA: O'Reilly).
- Mitchell,P., Waters,B., Vivero,C., Le,F., Donald,J., Tully,M., Campedelli,K., Lannfelt,L., Sokoloff,P., and Shine,J. (1993). Exclusion of close linkage of bipolar disorder to the dopamine D3 receptor gene in nine Australian pedigrees. J. Affect. Disord. *27*, 213-224.
- Mitchison,N.A. (2001). Polymorphism in regulatory gene sequences. Genome Biol. COMMENT2001.
- Miyoshi,K., Honda,A., Baba,K., Taniguchi,M., Oono,K., Fujita,T., Kuroda,S., Katayama,T., and Tohyama,M. (2003). Disrupted-In-Schizophrenia 1, a candidate gene for schizophrenia, participates in neurite outgrowth. Mol. Psychiatry *8*, 685-694.
- Mjellem,N., Lund,A., and Hole,K. (1993). Reduction of NMDA-induced behaviour after acute and chronic administration of desipramine in mice. Neuropharmacology *32*, 591-595.
- Moffatt,M.F. and Cookson,W.O. (1997). Tumour necrosis factor haplotypes and asthma. Hum. Mol. Genet. *6*, 551-554.
- Mohn,A.R., Gainetdinov,R.R., Caron,M.G., and Koller,B.H. (1999). Mice with reduced NMDA receptor expression display behaviors related to schizophrenia. Cell *98*, 427-436.

- Monaghan,G., Ryan,M., Seddon,R., Hume,R., and Burchell,B. (1996). Genetic variation in bilirubin UPD-glucuronosyltransferase gene promoter and Gilbert's syndrome. Lancet *347*, 578-581.
- Mooney,S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief. Bioinform. *6*, 44-56.
- Morel B.A. (1860). Traité des maladies mentales. (Paris: Masson).
- Morgenstern,B., Rinner,O., Abdeddaim,S., Haase,D., Mayer,K.F., Dress,A.W., and Mewes,H.W. (2002). Exon discovery by genomic sequence alignment. Bioinformatics *18*, 777-787.
- Morikawa,H., Imani,F., Khodakhah,K., and Williams,J.T. (2000). Inositol 1,4,5-triphosphate-evoked responses in midbrain dopamine neurons. J. Neurosci. *20*, RC103.
- Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S., and Cheung,V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743-747.
- Morris,D.W., McGhee,K.A., Schwaiger,S., Scully,P., Quinn,J., Meagher,D., Waddington,J.L., Gill,M., and Corvin,A.P. (2003). No evidence for association of the dysbindin gene [DTNBP1] with schizophrenia in an Irish population-based study. Schizophr. Res. *60*, 167-172.
- Morris,D.W., Rodgers,A., McGhee,K.A., Schwaiger,S., Scully,P., Quinn,J., Meagher,D., Waddington,J.L., Gill,M., and Corvin,A.P. (2004). Confirming RGS4 as a susceptibility gene for schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *125*, 50-53.
- Morrison,R.S., Kinoshita,Y., Johnson,M.D., Uo,T., Ho,J.T., McBee,J.K., Conrads,T.P., and Veenstra,T.D. (2002). Proteomic analysis in the neurosciences. Mol. Cell Proteomics. *1*, 553-560.
- Mortensen,O.V., Thomassen,M., Larsen,M.B., Whittemore,S.R., and Wiborg,O. (1999). Functional analysis of a novel human serotonin transporter gene promoter in immortalized raphe cells. Brain Res. Mol. Brain Res. *68*, 141-148.
- Morton,N.E. (1955). Sequential tests for the detection of linkage. Am. J. Hum. Genet. *7*, 277-318.
- Mothet,J.P., Parent,A.T., Wolosker,H., Brady,R.O., Jr., Linden,D.J., Ferris,C.D., Rogawski,M.A., and Snyder,S.H. (2000). D-serine is an endogenous ligand for the glycine site of the N-methyl-D-aspartate receptor. Proc. Natl. Acad. Sci. U. S. A. *97*, 4926-4931.
- Muglia,P., Petronis,A., Mundo,E., Lander,S., Cate,T., and Kennedy,J.L. (2002). Dopamine D4 receptor and tyrosine hydroxylase genes in bipolar disorder: evidence for a role of DRD4. Mol. Psychiatry *7*, 860-866.
- Muglia,P., Vicente,A.M., Verga,M., King,N., Macciardi,F., and Kennedy,J.L. (2003). Association between the BDNF gene and schizophrenia. Mol. Psychiatry *8*, 146-147.
- Muir,W.J., Gosden,C.M., Brookes,A.J., Fantes,J., Evans,K.L., Maguire,S.M., Stevenson,B., Boyle,S., Blackwood,D.H., and St Clair,D.M. (1995). Direct microdissection and microcloning of a translocation breakpoint region, t(1;11)(q42.2;q21), associated with schizophrenia. Cytogenet. Cell Genet. *70*, 35-40.
- Muir,W.J., Thomson,M.L., McKeon,P., Mynett-Johnson,L., Whitton,C., Evans,K.L., Porteous,D.J., and Blackwood,D.H. (2001). Markers close to the dopamine D5 receptor gene (DRD5) show significant association with schizophrenia but not bipolar disorder. Am. J. Med. Genet. *105*, 152-158.
- Mukai,J., Liu,H., Burt,R.A., Swor,D.E., Lai,W.S., Karayiorgou,M., and Gogos,J.A. (2004). Evidence that the gene encoding ZDHHC8 contributes to the risk of schizophrenia. Nat. Genet. *36*, 725-731.
- Muller,F. and Tora,L. (2004). The multicoloured world of promoter recognition complexes. EMBO J. *23*, 2-8.
- Munoz,D.G. and Feldman,H. (2000). Causes of Alzheimer's disease. CMAJ. *162*, 65-72.

- Muramatsu,T., Matsushita,S., Kanba,S., Higuchi,S., Manki,H., Suzuki,E., and Asai,M. (1997). Monoamine oxidase genes polymorphisms and mood disorder. Am. J. Med. Genet. *74*, 494-496.
- Murer,M.G., Yan,Q., and Raisman-Vozari,R. (2001). Brain-derived neurotrophic factor in the control human brain, and in Alzheimer's disease and Parkinson's disease. Prog. Neurobiol. *63*, 71-124.
- Murphy,K.C. (2002). Schizophrenia and velo-cardio-facial syndrome. Lancet *359*, 426-430.
- Murphy,K.C., Jones,L.A., and Owen,M.J. (1999). High rates of schizophrenia in adults with velo-cardio-facial syndrome. Arch. Gen. Psychiatry *56*, 940-945.
- Murray,C.J. and Lopez,A.D. (1997). Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. Lancet *349*, 1498-1504.
- Myers,J.S., Vincent,B.J., Udall,H., Watkins,W.S., Morrish,T.A., Kilroy,G.E., Swergold,G.D., Henke,J., Henke,L., Moran,J.V., Jorde,L.B., and Batzer,M.A. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am. J. Hum. Genet. *71*, 312-326.
- Nadler,J.V., Bray,S.D., and Evenson,D.A. (1992). Autoradiographic localization of proline uptake in excitatory hippocampal pathways. Hippocampus *2*, 269-278.
- Nakata,K., Takai,T., and Kaminuma,T. (1999). Development of the receptor database (RDB): application to the endocrine disruptor problem. Bioinformatics. *15*, 544-552.
- Nakayama,K., Fukamachi,S., Kimura,H., Koda,Y., Soemantri,A., and Ishida,T. (2002a). Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. J. Hum. Genet. *47*, 92-94.
- Nakayama,T., Soma,M., Mizutani,Y., Xinjuan,X., Honye,J., Kaneko,Y., Rahmutula,D., Aoi,N., Kosuge,K., Saito,S., Ozawa,Y., Kanmatsuse,K., and Kokubun,S. (2002b). A novel missense mutation of exon 3 in the type A human natriuretic peptide receptor gene: possible association with essential hypertension. Hypertens. Res. *25*, 395-401.
- Nakayama,T., Soma,M., Saito,S., Honye,J., Yajima,J., Rahmutula,D., Kaneko,Y., Sato,M., Uwabo,J., Aoi,N., Kosuge,K., Kunimoto,M., Kanmatsuse,K., and Kokubun,S. (2002c). Association of a novel single nucleotide polymorphism of the prostacyclin synthase gene with myocardial infarction. Am. Heart J. *143*, 797-801.
- Nakayama,T., Soma,M., Watanabe,Y., Hasimu,B., Sato,M., Aoi,N., Kosuge,K., Kanmatsuse,K., Kokubun,S., Marrow,J.D., and Oates,J.A. (2002d). Splicing mutation of the prostacyclin synthase gene in a family associated with hypertension. Biochem. Biophys. Res. Commun. *297*, 1135-1139.
- Nanko,S., Fukuda,R., Hattori,M., Sasaki,T., Dai,X.Y., Kanba,S., Kato,T., and Kazamatsuri,H. (1994). Linkage studies between affective disorder and dopamine D2, D3, and D4 receptor gene loci in four Japanese pedigrees. Psychiatry Res. *52*, 149-157.
- Nei,M. and Kumar,S. (2000). Molecular Evolution and Phylogenetics. (Oxford: Oxford University Press).
- Neves-Pereira,M., Mundo,E., Muglia,P., King,N., Macciardi,F., and Kennedy,J.L. (2002). The brain-derived neurotrophic factor gene confers susceptibility to bipolar disorder: evidence from a family-based association study. Am. J. Hum. Genet. *71*, 651-655.
- Neville,M.J., Johnstone,E.C., and Walton,R.T. (2004). Identification and characterization of ANKK1: a novel kinase gene closely linked to DRD2 on chromosome band 11q23.1. Hum. Mutat. *23*, 540-545.
- Newbury,D.F., Bonora,E., Lamb,J.A., Fisher,S.E., Lai,C.S., Baird,G., Jannoun,L., Slonims,V., Stott,C.M., Merricks,M.J., Bolton,P.F., Bailey,A.J., and Monaco,A.P. (2002). FOXP2 is not a major susceptibility gene for autism or specific language impairment. Am. J. Hum. Genet. *70*, 1318-1327.
- Ni,X., Trakalo,J.M., Mundo,E., Macciardi,F.M., Parikh,S., Lee,L., and Kennedy,J.L. (2002). Linkage disequilibrium between dopamine D1 receptor gene (DRD1) and bipolar disorder. Biol. Psychiatry *52*, 1144-1150.
- Nigumann,P., Redik,K., Matlik,K., and Speek,M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics *79*, 628-634.

- Nimpf,J. and Schneider,W.J. (2000). From cholesterol transport to signal transduction: low density lipoprotein receptor, very low density lipoprotein receptor, and apolipoprotein E receptor-2. Biochim. Biophys. Acta *1529*, 287-298.
- Nistico,L., Buzzetti,R., Pritchard,L.E., Van der,A.B., Giovannini,C., Bosi,E., Larrad,M.T., Rios,M.S., Chow,C.C., Cockram,C.S., Jacobs,K., Mijovic,C., Bain,S.C., Barnett,A.H., Vandewalle,C.L., Schuit,F., Gorus,F.K., Tosi,R., Pozzilli,P., and Todd,J.A. (1996). The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. Hum. Mol. Genet. *5*, 1075-1080.
- Noble,F., Wank,S.A., Crawley,J.N., Bradwejn,J., Seroogy,K.B., Hamon,M., and Roques,B.P. (1999). International Union of Pharmacology. XXI. Structure, distribution, and functions of cholecystokinin receptors. Pharmacol. Rev. *51*, 745-781.
- Nobrega,M.A., Ovcharenko,I., Afzal,V., and Rubin,E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.
- Nobrega,M.A. and Pennacchio,L.A. (2004). Comparative genomic analysis as a tool for biological discovery. J. Physiol *554*, 31-39.
- Nothen,M.M., Eggermann,K., Albus,M., Borrmann,M., Rietschel,M., Korner,J., Maier,W., Minges,J., Lichtermann,D., and Franzek,E. (1995). Association analysis of the monoamine oxidase A gene in bipolar affective disorder by using family-based internal controls. Am. J. Hum. Genet. *57*, 975-978.
- O'Donovan,M., Jones,I., and Craddock,N. (2003a). Anticipation and repeat expansion in bipolar disorder. Am. J. Med. Genet. C. Semin. Med. Genet. *123*, 10-17.
- O'Donovan,M.C., Guy,C., Craddock,N., Bowen,T., McKeon,P., Macedo,A., Maier,W., Wildenauer,D., Aschauer,H.N., Sorbi,S., Feldman,E., Mynett-Johnson,L., Claffey,E., Nacmias,B., Valente,J., Dourado,A., Grassi,E., Lenzinger,E., Heiden,A.M., Moorhead,S., Harrison,D., Williams,J., McGuffin,P., and Owen,M.J. (1996). Confirmation of association between expanded CAG/CTG repeats and both schizophrenia and bipolar disorder. Psychol. Med. *26*, 1145-1153.
- O'Donovan,M.C., Williams,N.M., and Owen,M.J. (2003b). Recent advances in the genetics of schizophrenia. Hum. Mol. Genet. *12 Spec No 2*, R125-R133.
- O'Reilly,R.L., Bogue,L., and Singh,S.M. (1994). Pharmacogenetic response to antidepressants in a multicase family with affective disorder. Biol. Psychiatry *36*, 467-471.
- Ogilvie,A.D., Battersby,S., Bubb,V.J., Fink,G., Harmar,A.J., Goodwim,G.M., and Smith,C.A. (1996). Polymorphism in serotonin transporter gene associated with susceptibility to major depression. Lancet *347*, 731-733.
- Ogiwara,I., Miya,M., Ohshima,K., and Okada,N. (2002). V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. Genome Res. *12*, 316-324.
- Ogura,Y., Bonen,D.K., Inohara,N., Nicolae,D.L., Chen,F.F., Ramos,R., Britton,H., Moran,T., Karaliuskas,R., Duerr,R.H., Achkar,J.P., Brant,S.R., Bayless,T.M., Kirschner,B.S., Hanauer,S.B., Nunez,G., and Cho,J.H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature *411*, 603-606.
- Ohara,K., Nagai,M., Tani,K., Nakamura,Y., Ino,A., and Ohara,K. (1998). Functional polymorphism of -141C Ins/Del in the dopamine D2 receptor gene promoter and schizophrenia. Psychiatry Res. *81*, 117-123.
- Ohler,U. and Niemann,H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet. *17*, 56-60.
- Ohtsuki,T., Tanaka,S., Ishiguro,H., Noguchi,E., Arinami,T., Tanabe,E., Yara,K., Okubo,T., Takahashi,S., Matsuura,M., Sakai,T., Muto,M., Kojima,T., Matsushima,E., Toru,M., and Inada,T. (2004). Failure to find association between PRODH deletion and schizophrenia. Schizophr. Res. *67*, 111-113.
- Okada,N. (1991). SINEs. Curr. Opin. Genet. Dev. *1*, 498-504.
- Okubo,T., Harada,S., Higuchi,S., and Matsushita,S. (2002). Investigation of quantitative trait loci in the CCKAR gene with susceptibility to alcoholism. Alcohol Clin. Exp. Res. *26*, 2S-5S.

- Okuyama,Y., Ishiguro,H., Nankai,M., Shibuya,H., Watanabe,A., and Arinami,T. (2000). Identification of a polymorphism in the promoter region of DRD4 associated with the human novelty seeking personality trait. Mol. Psychiatry *5*, 64-69.
- Okuyama,Y., Ishiguro,H., Toru,M., and Arinami,T. (1999). A genetic polymorphism in the promoter region of DRD4 associated with expression and schizophrenia. Biochem. Biophys. Res. Commun. *258*, 292-295.
- Oleksiak,M.F., Churchill,G.A., and Crawford,D.L. (2002). Variation in gene expression within and among natural populations. Nat. Genet. *32*, 261-266.
- Oliveri,R.L., Annesi,G., Zappia,M., Civitelli,D., Montesanti,R., Branca,D., Nicoletti,G., Spadafora,P., Pasqua,A.A., Cittadella,R., Andreoli,V., Gambardella,A., Aguglia,U., and Quattrone,A. (1999). Dopamine D2 receptor gene polymorphism and the risk of levodopa-induced dyskinesias in PD. Neurology *53*, 1425-1430.
- Olson,M.V. and Varki,A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. Nat. Rev. Genet. *4*, 20-28.
- Orfali,R. and Harkey,D. (1998). Client/server programming with Java and CORBA. (New York: J. Wiley & Sons).
- Orphanides,G. and Reinberg,D. (2002). A unified theory of gene expression. Cell *108*, 439-451.
- Ostertag,E.M. and Kazazian,H.H., Jr. (2001). Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. *35*, 501-538.
- Owen,M.J. and Cardno,A.G. (1999). Psychiatric genetics: progress, problems, and potential. Lancet *354 Suppl 1*, SI11-SI14.
- Owen,M.J., Williams,N.M., and O'Donovan,M.C. (2004). Dysbindin-1 and schizophrenia: from genetics to neuropathology. J. Clin. Invest. *113*, 1255-1257.
- Ozaki,K., Ohnishi,Y., Iida,A., Sekine,A., Yamada,R., Tsunoda,T., Sato,H., Sato,H., Hori,M., Nakamura,Y., and Tanaka,T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. *32*, 650-654.
- Ozaki,N., Goldman,D., Kaye,W.H., Plotnicov,K., Greenberg,B.D., Lappalainen,J., Rudnick,G., and Murphy,D.L. (2003). Serotonin transporter missense mutation associated with a complex neuropsychiatric phenotype. Mol. Psychiatry *8*, 895, 933-895, 936.
- Ozeki,Y., Tomoda,T., Kleiderlein,J., Kamiya,A., Bord,L., Fujii,K., Okawa,M., Yamada,N., Hatten,M.E., Snyder,S.H., Ross,C.A., and Sawa,A. (2003). Disrupted-in-Schizophrenia-1 (DISC-1): mutant truncation prevents binding to NudE-like (NUDEL) and inhibits neurite outgrowth. Proc. Natl. Acad. Sci. U. S. A. *100*, 289-294.
- Page,N.M., Butlin,D.J., Lomthaisong,K., and Lowry,P.J. (2001). The human apolipoprotein L gene cluster: identification, classification, and sites of distribution. Genomics *74*, 71-78.
- Papolos,D.F., Veit,S., Faedda,G.L., Saito,T., and Lachman,H.M. (1998). Ultra-ultra rapid cycling bipolar disorder is associated with the low activity catecholamine-O-methyltransferase allele. Mol. Psychiatry *3*, 346-349.
- Pare,C.M. and Mack,J.W. (1971). Differentiation of two genetically specific types of depression by the response to antidepressant drugs. J. Med. Genet. *8*, 306-309.
- Pastinen,T., Sladek,R., Gurd,S., Sammak,A., Ge,B., Lepage,P., Lavergne,K., Villeneuve,A., Gaudin,T., Brandstrom,H., Beck,A., Verner,A., Kingsley,J., Harmsen,E., Labuda,D., Morgan,K., Vohl,M.C., Naumova,A.K., Sinnett,D., and Hudson,T.J. (2004). A survey of genetic and epigenetic variation affecting human gene expression. Physiol. Genomics *16*, 184-193.
- Paterson,D. and Nordberg,A. (2000). Neuronal nicotinic receptors in the human brain. Prog. Neurobiol. *61*, 75-111.
- Patil,N., Berno,A.J., Hinds,D.A., Barrett,W.A., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C., McDonough,D.P., Nguyen,B.T., Norris,M.C., Sheehan,J.B., Shen,N., Stern,D., Stokowski,R.P., Thomas,D.J., Trulson,M.O., Vyas,K.R., Frazer,K.A., Fodor,S.P., and Cox,D.R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science *294*, 1719-1723.

- Patrinos,G.P., Giardine,B., Riemer,C., Miller,W., Chui,D.H., Anagnou,N.P., Wajcman,H., and Hardison,R.C. (2004). Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res. *32*, D537-D541.
- Pavesi,G., Mauri,G., and Pesole,G. (2004). In silico representation and discovery of transcription factor binding sites. Brief. Bioinform. *5*, 217-236.
- Payton,A., Holmes,J., Barrett,J.H., Hever,T., Fitzpatrick,H., Trumper,A.L., Harrington,R., McGuffin,P., O'Donovan,M., Owen,M., Ollier,W., Worthington,J., and Thapar,A. (2001). Examining for association between candidate gene polymorphisms in the dopamine pathway and attention-deficit hyperactivity disorder: a family-based study. Am. J. Med. Genet. *105*, 464-470.
- Pearson,E.S., D'Agostino,R.B., and Bowman,K.O. (1977). Tests for departure from normality: comparison of powers. Biometrika 231-246.
- Peltonen,L. and McKusick,V.A. (2001). Genomics and medicine. Dissecting human disease in the postgenomic era. Science *291*, 1224-1229.
- Pennacchio,L.A. and Rubin,E.M. (2001). Genomic strategies to identify mammalian regulatory sequences. Nat. Rev. Genet. *2*, 100-109.
- Perier,R.C., Praz,V., Junier,T., Bonnard,C., and Bucher,P. (2000). The eukaryotic promoter database (EPD). Nucleic Acids Res. *28*, 302-303.
- Petronis,A. (2001). Human morbid genetics revisited: relevance of epigenetics. Trends Genet. *17*, 142-146.
- Petronis,A. (2003). Epigenetics and bipolar disorder: new opportunities and challenges. Am. J. Med. Genet. C. Semin. Med. Genet. *123*, 65-75.
- Petronis,A., Gottesman,I.I., Crow,T.J., DeLisi,L.E., Klar,A.J., Macciardi,F., McInnis,M.G., McMahon,F.J., Paterson,A.D., Skuse,D., and Sutherland,G.R. (2000). Psychiatric epigenetics: a new focus for the new century. Mol. Psychiatry *5*, 342-346.
- Petronis,A., Gottesman,I.I., Kan,P., Kennedy,J.L., Basile,V.S., Paterson,A.D., and Popendikyte,V. (2003). Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance? Schizophr. Bull. *29*, 169-178.
- Petronis,A. and Petroniene,R. (2000). Epigenetics of inflammatory bowel disease. Gut *47*, 302-306.
- Phang,J.M., Hu,C.-A., and Valle,D. (2001). Disorders of proline and hydroxyproline metabolism. In The metabolic and molecular bases of inherited disease, C.R.Scriver, A.L.Beaudet, W.S.Sly, and D.Valle, eds. (New York: McGraw Hill), pp. 1821-1838.
- Pickard,B.S., Millar,J.K., Porteous,D.J., Muir,W.J., and Blackwood,D.H. (2005). Cytogenetics and gene discovery in psychiatric disorders. Pharmacogenomics *5*, 81-88.
- Picketts,D.J., Mueller,C.R., and Lillicrap,D. (1994). Transcriptional control of the factor IX gene: analysis of five cis-acting elements and the deleterious effects of naturally occurring hemophilia B Leyden mutations. Blood *84*, 2992-3000.
- Pinel (1798). Nosographie philosophique ou méthode de l'analyse appliquée à la medecine. (Paris: Brosson).
- Ping,H.X. and Shepard,P.D. (1999). Blockade of SK-type Ca2+-activated K+ channels uncovers a Ca2+-dependent slow afterdepolarization in nigral dopamine neurons. J. Neurophysiol. *81*, 977-984.
- Pisegna,J.R., De Weerth,A., Huppi,K., and Wank,S.A. (1992). Molecular cloning of the human brain and gastric cholecystokinin receptor: structure, functional expression and chromosomal localization. Biochem. Biophys. Res. Commun. *189*, 296-303.
- Placzek,M.R., Misbahuddin,A., Chaudhuri,K.R., Wood,N.W., Bhatia,K.P., and Warner,T.T. (2001). Cervical dystonia is associated with a polymorphism in the dopamine (D5) receptor gene. J. Neurol. Neurosurg. Psychiatry *71*, 262-264.
- Plyte,S.E., Hughes,K., Nikolakaki,E., Pulverer,B.J., and Woodgett,J.R. (1992). Glycogen synthase kinase-3: functions in oncogenesis and development. Biochim. Biophys. Acta *1114*, 147-162.

- Pollet,N., Schmidt,H.A., Gawantka,V., Vingron,M., and Niehrs,C. (2000). Axeldb: a Xenopus laevis database focusing on gene expression. Nucleic Acids Res. *28*, 139-140.
- Pollock,B.G., Mulsant,B.H., Nebes,R., Kirshner,M.A., Begley,A.E., Mazumdar,S., and Reynolds,C.F., III (1998). Serum anticholinergicity in elderly depressed patients treated with paroxetine or nortriptyline. Am. J. Psychiatry *155*, 1110-1112.
- Polymeropoulos,M.H. and Schaffer,A.A. (1996). Scanning the genome with 1772 microsatellite markers in search of a bipolar disorder susceptibility gene. Mol. Psychiatry *1*, 404-407.
- Porteous,D.J., Evans,K.L., Millar,J.K., Pickard,B.S., Thomson,P.A., James,R., MacGregor,S., Wray,N.R., Visscher,P.M., Muir,W.J., and Blackwood,D.H. (2003). Genetics of schizophrenia and bipolar affective disorder: strategies to identify candidate genes. Cold Spring Harb. Symp. Quant. Biol. *68*, 383-394.
- Potash,J.B. and DePaulo,J.R. (2000). Searching high and low: a review of the genetics of bipolar disorder. Bipolar. Disord. *2*, 8-26.
- Powers,D.A. (1991). Evolutionary genetics of fish. Adv. Genet. *29*, 119-228.
- Proudfoot,N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. Curr. Opin. Cell Biol. *16*, 272-278.
- Proudfoot,N.J., Furger,A., and Dye,M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.
- Pulver,A.E., Nestadt,G., Goldberg,R., Shprintzen,R.J., Lamacz,M., Wolyniec,P.S., Morrow,B., Karayiorgou,M., Antonarakis,S.E., and Housman,D. (1994). Psychotic illness in patients diagnosed with velo-cardio-facial syndrome and their relatives. J. Nerv. Ment. Dis. *182*, 476-478.
- Quandt,K., Frech,K., Karas,H., Wingender,E., and Werner,T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. *23*, 4878-4884.
- Quentin,Y. and Fichant,G. (2000). ABCdb: an ABC transporter database. J. Mol. Microbiol. Biotechnol. *2*, 501-504.
- Ramakrishnan,V. (2002). Ribosome structure and the mechanism of translation. Cell *108*, 557-572.
- Ranade,K., Chang,M.S., Ting,C.T., Pei,D., Hsiao,C.F., Olivier,M., Pesich,R., Hebert,J., Chen,Y.D., Dzau,V.J., Curb,D., Olshen,R., Risch,N., Cox,D.R., and Botstein,D. (2001). High-throughput genotyping with single nucleotide polymorphisms. Genome Res. *11*, 1262-1268.
- Rao,D., Jonsson,E.G., Paus,S., Ganguli,R., Nothen,M., and Nimgaonkar,V.L. (1998). Schizophrenia and the serotonin transporter gene. Psychiatr. Genet. *8*, 207-212.
- Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.
- Rave-Harel,N., Kerem,E., Nissim-Rafinia,M., Madjar,I., Goshen,R., Augarten,A., Rahat,A., Hurwitz,A., Darvasi,A., and Kerem,B. (1997). The molecular basis of partial penetrance of splicing mutations in cystic fibrosis. Am. J. Hum. Genet. *60*, 87-94.
- Reed,R. and Hurt,E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell *108*, 523-531.
- Rehfeld,J.F. (2004). Clinical endocrinology and metabolism. Cholecystokinin. Best. Pract. Res. Clin. Endocrinol. Metab. *18*, 569-586.
- Reich,D.E. and Lander,E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502-510.
- Reich,D.E., Schaffner,S.F., Daly,M.J., McVean,G., Mullikin,J.C., Higgins,J.M., Richter,D.J., Lander,E.S., and Altshuler,D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. *32*, 135-142.
- Renick,S.E., Kleven,D.T., Chan,J., Stenius,K., Milner,T.A., Pickel,V.M., and Fremeau,R.T., Jr. (1999). The mammalian brain high-affinity L-proline transporter is enriched preferentially

in synaptic vesicles in a subpopulation of excitatory nerve terminals in rat forebrain. J. Neurosci. *19*, 21-33.

- Rice,D.P. (1999). The economic impact of schizophrenia. J. Clin. Psychiatry *60 Suppl 1*, 4-6.
- Rickards,H. (2005). Depression in neurological disorders: Parkinson's disease, multiple sclerosis, and stroke. J. Neurol. Neurosurg. Psychiatry *76 Suppl 1*, i48-i52.
- Rifkin,S.A., Kim,J., and White,K.P. (2003). Evolution of gene expression in the Drosophila melanogaster subgroup. Nat. Genet. *33*, 138-144.
- Rimini,R., Rimland,J.M., and Terstappen,G.C. (2000). Quantitative expression analysis of the small conductance calcium-activated potassium channels, SK1, SK2 and SK3, in human brain. Brain Res. Mol. Brain Res. *85*, 218-220.
- Rioux,J.D., Silverberg,M.S., Daly,M.J., Steinhart,A.H., McLeod,R.S., Griffiths,A.M., Green,T., Brettin,T.S., Stone,V., Bull,S.B., Bitton,A., Williams,C.N., Greenberg,G.R., Cohen,Z., Lander,E.S., Hudson,T.J., and Siminovitch,K.A. (2000). Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. Am. J. Hum. Genet. *66*, 1863-1870.
- Risch,N. and Botstein,D. (1996). A manic depressive history. Nat. Genet. *12*, 351-353.
- Risch,N. and Merikangas,K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.
- Risch,N. and Teng,J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res. *8*, 1273-1288.
- Risch,N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847-856.
- Ritsner,M., Amir,S., Koronyo-Hamaoui,M., Gak,E., Ziv,H., Halperin,T., Kitain,L., and Navon,R. (2003). Association study of CAG repeats in the KCNN3 gene in Israeli patients with major psychosis. Psychiatr. Genet. *13*, 143-150.
- Ritsner,M., Modai,I., Ziv,H., Amir,S., Halperin,T., Weizman,A., and Navon,R. (2002). An association of CAG repeats at the KCNN3 locus with symptom dimensions of schizophrenia. Biol. Psychiatry *51*, 788-794.
- Roberts,G.W. (1990). Schizophrenia: the cellular biology of a functional psychosis. Trends Neurosci. *13*, 207-211.
- Rockman,M.V., Hahn,M.W., Soranzo,N., Goldstein,D.B., and Wray,G.A. (2003). Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. *13*, 2118-2123.
- Rockman,M.V. and Wray,G.A. (2002). Abundant raw material for cis-regulatory evolution in humans. Mol. Biol. Evol. *19*, 1991-2004.
- Rogers,G., Joyce,P., Mulder,R., Sellman,D., Miller,A., Allington,M., Olds,R., Wells,E., and Kennedy,M. (2004). Association of a duplicated repeat polymorphism in the 5'-untranslated region of the DRD4 gene with novelty seeking. Am. J. Med. Genet. B Neuropsychiatr. Genet. *126*, 95-98.
- Rohrbough,J. and Broadie,K. (2005). Lipid regulation of the synaptic vesicle cycle. Nat. Rev. Neurosci. *6*, 139-150.
- Rohrmeier,T., Putzhammer,A., Schoeler,A., Sartor,H., Dallinger,P., Nothen,M.M., Propping,P., Knapp,M., Albus,M., Borrmann,M., Knothe,K., Kreiner,R., Franzek,E., Lichtermann,D., Rietschel,M., Maier,W., Klein,H.E., and Eichhammer,P. (1999). hSKCa3: no association of the polymorphic CAG repeat with bipolar affective disorder and schizophrenia. Psychiatr. Genet. *9*, 169-175.
- Ronai,Z., Szekely,A., Nemoda,Z., Lakatos,K., Gervai,J., Staub,M., and Sasvari-Szekely,M. (2001). Association between Novelty Seeking and the -521 C/T polymorphism in the promoter region of the DRD4 gene. Mol. Psychiatry *6*, 35-38.
- Ronshaugen,M., McGinnis,N., and McGinnis,W. (2002). Hox protein mutation and macroevolution of the insect body plan. Nature *415*, 914-917.

- Roses,A.D. (1994). Apolipoprotein E affects the rate of Alzheimer disease expression: beta-amyloid burden is a secondary consequence dependent on APOE genotype and duration of disease. J. Neuropathol. Exp. Neurol. *53*, 429-437.
- Roses,A.D. (2000). Pharmacogenetics and the practice of medicine. Nature *405*, 857-865.
- Ross,E.M. and Wilkie,T.M. (2000). GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins. Annu. Rev. Biochem. *69*, 795-827.
- Roth,F.P., Hughes,J.D., Estep,P.W., and Church,G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. *16*, 939-945.
- Rotondo,A., Mazzanti,C., Dell'Osso,L., Rucci,P., Sullivan,P., Bouanani,S., Gonnelli,C., Goldman,D., and Cassano,G.B. (2002). Catechol o-methyltransferase, serotonin transporter, and tryptophan hydroxylase gene polymorphisms in bipolar disorder patients with and without comorbid panic disorder. Am. J. Psychiatry *159*, 23-29.
- Rotzinger,S. and Vaccarino,F.J. (2003). Cholecystokinin receptor subtypes: role in the modulation of anxiety-related and reward-related behaviours in animal models. J. Psychiatry Neurosci. *28*, 171-181.
- Roulet,E., Bucher,P., Schneider,R., Wingender,E., Dusserre,Y., Werner,T., and Mermod,N. (2000). Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. J. Mol. Biol. *297*, 833-848.
- Royston,J.P. (1982). An extension of the Shapiro and Wilk's W test for normality for large samples. Applied statistics *31*, 115-124.
- Rubinstein,M., Phillips,T.J., Bunzow,J.R., Falzone,T.L., Dziewczapolski,G., Zhang,G., Fang,Y., Larson,J.L., McDougall,J.A., Chester,J.A., Saez,C., Pugsley,T.A., Gershanik,O., Low,M.J., and Grandy,D.K. (1997). Mice lacking dopamine D4 receptors are supersensitive to ethanol, cocaine, and methamphetamine. Cell *90*, 991-1001.
- Rubinsztein,D.C., Leggo,J., Goodburn,S., Walsh,C., Jain,S., and Paykel,E.S. (1996). Genetic association between monoamine oxidase A microsatellite and RFLP alleles and bipolar affective disorder: analysis and meta-analysis. Hum. Mol. Genet. *5*, 779-782.
- Rybakowski,J.K., Borkowska,A., Czerski,P.M., and Hauser,J. (2001). Dopamine D3 receptor (DRD3) gene polymorphism is associated with the intensity of eye movement disturbances in schizophrenic patients and healthy subjects. Mol. Psychiatry *6*, 718-724.
- Sabol,S.Z., Hu,S., and Hamer,D. (1998). A functional polymorphism in the monoamine oxidase A gene promoter. Hum. Genet. *103*, 273-279.
- Sah,P. and Davies,P. (2000). Calcium-activated potassium currents in mammalian neurons. Clin. Exp. Pharmacol. Physiol. *27*, 657-663.
- Saito,S., Ikeda,M., Iwata,N., Suzuki,T., Kitajima,T., Yamanouchi,Y., Kinoshita,Y., Takahashi,N., Inada,T., and Ozaki,N. (2005). No association was found between a functional SNP in ZDHHC8 and schizophrenia in a Japanese case-control population. Neurosci. Lett. *374*, 21-24.
- Saito,T., Guan,F., Papolos,D.F., Lau,S., Klein,M., Fann,C.S., and Lachman,H.M. (2001). Mutation analysis of SYNJ1: a possible candidate gene for chromosome 21q22-linked bipolar disorder. Mol. Psychiatry *6*, 387-395.
- Saito,T., Stopkova,P., Diaz,L., Papolos,D.F., Boussemart,L., and Lachman,H.M. (2003). Polymorphism screening of PIK4CA: possible candidate gene for chromosome 22q11-linked psychiatric disorders. Am. J. Med. Genet. B Neuropsychiatr. Genet. *116*, 77-83.
- Sakolsky,D.J. and Ashby,B. (2001). Patterns of cyclic AMP formation by coexpressed D1 and D2L dopamine receptors in HEK 293 cells. Receptors. Channels *7*, 479-489.
- Saleem,Q., Sreevidya,V.S., Sudhir,J., Savithri,J.V., Gowda,Y., Rao,C., Benegal,V., Majumder,P.P., Anand,A., Brahmachari,S.K., and Jain,S. (2000). Association analysis of CAG repeats at the KCNN3 locus in Indian patients with bipolar disorder and schizophrenia. Am. J. Med. Genet. *96*, 744-748.

- Samuelson,L.C., Wiebauer,K., Snow,C.M., and Meisler,M.H. (1990). Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol. Cell Biol. *10*, 2513-2520.
- Sandbrink,R., Hartmann,T., Masters,C.L., and Beyreuther,K. (1996). Genes contributing to Alzheimer's disease. Mol. Psychiatry *1*, 27-40.
- Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W., and Lenhard,B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32 Database issue*, D91-D94.
- Sanfilipo,M., Lafargue,T., Rusinek,H., Arena,L., Loneragan,C., Lautin,A., Feiner,D., Rotrosen,J., and Wolkin,A. (2000). Volumetric measure of the frontal and temporal lobe regions in schizophrenia: relationship to negative symptoms. Arch. Gen. Psychiatry *57*, 471-480.
- Sasaki,T., Hattori,M., Sakai,T., Kato,T., Kunugi,H., Hirose,T., and Nanko,S. (1998). The monoamine oxidase-A gene and major psychosis in Japanese subjects. Biol. Psychiatry *44*, 922-924.
- Saunders,N.J., Peden,J.F., Hood,D.W., and Moxon,E.R. (1998). Simple sequence repeats in the Helicobacter pylori genome. Mol. Microbiol. *27*, 1091-1098.
- Schadt,E.E., Monks,S.A., Drake,T.A., Lusis,A.J., Che,N., Colinayo,V., Ruff,T.G., Milligan,S.B., Lamb,J.R., Cavet,G., Linsley,P.S., Mao,M., Stoughton,R.B., and Friend,S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297-302.
- Schiffer,H.H. (2002). Glutamate receptor genes: susceptibility factors in schizophrenia and depressive disorders? Mol. Neurobiol. *25*, 191-212.
- Scholl,R. (2002). Der papyrus Ebers. Die grosste buchrolle zur heilkunde altagyptens.
- Schulze,T.G., Buervenich,S., Badner,J.A., Steele,C.J., Detera-Wadleigh,S.D., Dick,D., Foroud,T., Cox,N.J., MacKinnon,D.F., Potash,J.B., Berrettini,W.H., Byerley,W., Coryell,W., DePaulo,J.R., Jr., Gershon,E.S., Kelsoe,J.R., McInnis,M.G., Murphy,D.L., Reich,T., Scheftner,W., Nurnberger,J.I., Jr., and McMahon,F.J. (2004). Loci on chromosomes 6q and 6p interact to increase susceptibility to bipolar affective disorder in the national institute of mental health genetics initiative pedigrees. Biol. Psychiatry *56*, 18-23.
- Schulze,T.G., Muller,D.J., Krauss,H., Scherk,H., Ohlraun,S., Syagailo,Y.V., Windemuth,C., Neidt,H., Grassle,M., Papassotiropoulos,A., Heun,R., Nothen,M.M., Maier,W., Lesch,K.P., and Rietschel,M. (2000). Association between a functional polymorphism in the monoamine oxidase A gene promoter and major depressive disorder. Am. J. Med. Genet. *96*, 801-803.
- Schumacher,J., Jamra,R.A., Freudenberg,J., Becker,T., Ohlraun,S., Otte,A.C., Tullius,M., Kovalenko,S., Bogaert,A.V., Maier,W., Rietschel,M., Propping,P., Nothen,M.M., and Cichon,S. (2004). Examination of G72 and D-amino-acid oxidase as genetic risk factors for schizophrenia and bipolar affective disorder. Mol. Psychiatry *9*, 203-207.
- Schwab,S.G., Knapp,M., Mondabon,S., Hallmayer,J., Borrmann-Hassenbach,M., Albus,M., Lerer,B., Rietschel,M., Trixler,M., Maier,W., and Wildenauer,D.B. (2003). Support for association of schizophrenia with genetic variation in the 6p22.3 gene, dysbindin, in sib-pair families with linkage and in an additional sample of triad families. Am. J. Hum. Genet. *72*, 185-190.
- Schwartz,S., Elnitski,L., Li,M., Weirauch,M., Riemer,C., Smit,A., Green,E.D., Hardison,R.C., and Miller,W. (2003a). MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res. *31*, 3518-3524.
- Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D., and Miller,W. (2003b). Human-mouse alignments with BLASTZ. Genome Res. *13*, 103-107.
- Schwartz,S., Miller,W., Yang,C.M., and Hardison,R.C. (1991). Software tools for analyzing pairwise alignments of long sequences. Nucleic Acids Res. *19*, 4663-4667.
- Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R., and Miller,W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. Genome Res. *10*, 577-586.
- Schweisguth,F. (2004). Notch signaling activity. Curr. Biol. *14*, R129-R138.

- Searle,S.M., Gilbert,J., Iyer,V., and Clamp,M. (2004). The otter annotation system. Genome Res. *14*, 963-970.
- Sedler,M.J. (1983). Falret's discovery: the origin of the concept of bipolar affective illness. Translated by M. J. Sedler and Eric C. Dessain. Am. J. Psychiatry *140*, 1127-1133.
- Seelan,R.S., Parthasarathy,L.K., and Parthasarathy,R.N. (2004). Lithium modulation of the human inositol monophosphatase 2 (IMPA2) promoter. Biochem. Biophys. Res. Commun. *324*, 1370-1378.
- Seeman,P. (1987). Dopamine receptors and the dopamine hypothesis of schizophrenia. Synapse *1*, 133-152.
- Semple,C.A., Morris,S.W., Porteous,D.J., and Evans,K.L. (2000). In silico identification of transcripts and SNPs from a region of 4p linked with bipolar affective disorder. Bioinformatics *16*, 735-738.
- Semple,C.A., Taylor,M.S., and Ballereau,S. (2001). The meso-genomic era. Genome Biol. *2*, REPORTS4015.
- Serretti,A., Cristina,S., Lilli,R., Cusin,C., Lattuada,E., Lorenzi,C., Corradi,B., Grieco,G., Costa,A., Santorelli,F., Barale,F., Nappi,G., and Smeraldi,E. (2002). Family-based association study of 5-HTTLPR, TPH, MAO-A, and DRD4 polymorphisms in mood disorders. Am. J. Med. Genet. *114*, 361-369.
- Serretti,A., Cusin,C., Rossini,D., Artioli,P., Dotoli,D., and Zanardi,R. (2004). Further evidence of a combined effect of SERTPR and TPH on SSRIs response in mood disorders. Am. J. Med. Genet. B Neuropsychiatr. Genet. *129*, 36-40.
- Serretti,A., Lattuada,E., Lorenzi,C., Lilli,R., and Smeraldi,E. (2000a). Dopamine receptor D2 Ser/Cys 311 variant is associated with delusion and disorganization symptomatology in major psychoses. Mol. Psychiatry *5*, 270-274.
- Serretti,A., Macciardi,F., Catalano,M., Bellodi,L., and Smeraldi,E. (1999). Genetic variants of dopamine receptor D4 and psychopathology. Schizophr. Bull. *25*, 609-618.
- Serretti,A., Macciardi,F., Cusin,C., Lattuada,E., Lilli,R., and Smeraldi,E. (1998). Dopamine receptor D4 gene is associated with delusional symptomatology in mood disorders. Psychiatry Res. *80*, 129-136.
- Serretti,A., Macciardi,F., Cusin,C., Lattuada,E., Souery,D., Lipp,O., Mahieu,B., Van Broeckhoven,C., Blackwood,D., Muir,W., Aschauer,H.N., Heiden,A.M., Ackenheil,M., Fuchshuber,S., Raeymaekers,P., Verheyen,G., Kaneva,R., Jablensky,A., Papadimitriou,G.N., Dikeos,D.G., Stefanis,C.N., Smeraldi,E., and Mendlewicz,J. (2000b). Linkage of mood disorders with D2, D3 and TH genes: a multicenter study. J. Affect. Disord. *58*, 51-61.
- Sestan,N., Artavanis-Tsakonas,S., and Rakic,P. (1999). Contact-dependent inhibition of cortical neurite growth mediated by notch signaling. Science *286*, 741-746.
- Severino,G., Congiu,D., Serreli,C., De Lisa,R., Chillotti,C., Del Zompo,M., and Piccardi,M.P. (2005). A48G polymorphism in the D1 receptor genes associated with bipolar I disorder. Am. J. Med. Genet. B Neuropsychiatr. Genet. *134*, 37-38.
- Shaikh,S., Ball,D., Craddock,N., Castle,D., Hunt,N., Mant,R., Owen,M., Collier,D., and Gill,M. (1993). The dopamine D3 receptor gene: no association with bipolar affective disorder. J. Med. Genet. *30*, 308-309.
- Shaikh,S., Collier,D., Arranz,M., Ball,D., Gill,M., and Kerwin,R. (1994). DRD2 Ser311/Cys311 polymorphism in schizophrenia. Lancet *343*, 1045-1046.
- Shaltiel,G., Shamir,A., Shapiro,J., Ding,D., Dalton,E., Bialer,M., Harwood,A.J., Belmaker,R.H., Greenberg,M.L., and Agam,G. (2004). Valproate decreases inositol biosynthesis. Biol. Psychiatry *56*, 868-874.
- Shankar,R., Grover,D., Brahmachari,S.K., and Mukerji,M. (2004). Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. BMC. Evol. Biol. *4*, 37.
- Shapiro,M.D., Marks,M.E., Peichel,C.L., Blackman,B.K., Nereng,K.S., Jonsson,B., Schluter,D., and Kingsley,D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature *428*, 717-723.

- Shapiro,S.S. and Wilk,M.B. (1965). An analysis of variance test for normality. Biometrika *52*, 591-611.
- Shashikant,C.S., Kim,C.B., Borbely,M.A., Wang,W.C., and Ruddle,F.H. (1998). Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a cis-acting element. Proc. Natl. Acad. Sci. U. S. A *95*, 15446-15451.
- Shaw,S.H., Kelly,M., Smith,A.B., Shields,G., Hopkins,P.J., Loftus,J., Laval,S.H., Vita,A., De Hert,M., Cardon,L.R., Crow,T.J., Sherrington,R., and DeLisi,L.E. (1998). A genome-wide search for schizophrenia susceptibility genes. Am. J. Med. Genet. *81*, 364-376.
- Shewchuk,B.M., Cooke,N.E., and Liebhaber,S.A. (2001). The human growth hormone locus control region mediates long-distance transcriptional activation independent of nuclear matrix attachment regions. Nucleic Acids Res. *29*, 3356-3361.
- Shield,A.J., Thomae,B.A., Eckloff,B.W., Wieben,E.D., and Weinshilboum,R.M. (2004). Human catechol O-methyltransferase genetic variation: gene resequencing and functional characterization of variant allozymes. Mol. Psychiatry *9*, 151-160.
- Shifman,S., Bronstein,M., Sternfeld,M., Pisante,A., Weizman,A., Reznik,I., Spivak,B., Grisaru,N., Karp,L., Schiffer,R., Kotler,M., Strous,R.D., Swartz-Vanetik,M., Knobler,H.Y., Shinar,E., Yakir,B., Zak,N.B., and Darvasi,A. (2004). COMT: a common susceptibility gene in bipolar disorder and schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *128*, 61-64.
- Shifman,S., Bronstein,M., Sternfeld,M., Pisante-Shalom,A., Lev-Lehman,E., Weizman,A., Reznik,I., Spivak,B., Grisaru,N., Karp,L., Schiffer,R., Kotler,M., Strous,R.D., Swartz-Vanetik,M., Knobler,H.Y., Shinar,E., Beckmann,J.S., Yakir,B., Risch,N., Zak,N.B., and Darvasi,A. (2002). A highly significant association between a COMT haplotype and schizophrenia. Am. J. Hum. Genet. *71*, 1296-1302.
- Shilatifard,A. (2004). Transcriptional elongation control by RNA polymerase II: a new frontier. Biochim. Biophys. Acta *1677*, 79-86.
- Shink,E., Morissette,J., Sherrington,R., and Barden,N. (2005). A genome-wide scan points to a susceptibility locus for bipolar disorder on chromosome 12. Mol. Psychiatry *10*, 545-552.
- Sibbing,D., Asmus,F., Konig,I.R., Tezenas du,M.S., Vidailhet,M., Sangla,S., Oertel,W.H., Brice,A., Ziegler,A., Gasser,T., and Bandmann,O. (2003). Candidate gene studies in focal dystonia. Neurology *61*, 1097-1101.
- Simon,G.E. (2003). Social and economic burden of mood disorders. Biol. Psychiatry *54*, 208-215.
- Simonic,I., Nyholt,D.R., Gericke,G.S., Gordon,D., Matsumoto,N., Ledbetter,D.H., Ott,J., and Weber,J.L. (2001). Further evidence for linkage of Gilles de la Tourette syndrome (GTS) susceptibility loci on chromosomes 2p11, 8q22 and 11q23-24 in South African Afrikaners. Am. J. Med. Genet. *105*, 163-167.
- Sklar,P. (2002). Linkage analysis in psychiatric disorders: the emerging picture. Annu. Rev. Genomics Hum. Genet. *3*, 371-413.
- Sklar,P., Schwab,S.G., Williams,N.M., Daly,M., Schaffner,S., Maier,W., Albus,M., Trixler,M., Eichhammer,P., Lerer,B., Hallmayer,J., Norton,N., Williams,H., Zammit,S., Cardno,A.G., Jones,S., McCarthy,G., Milanova,V., Kirov,G., O'Donovan,M.C., Lander,E.S., Owen,M.J., and Wildenauer,D.B. (2001). Association analysis of NOTCH4 loci in schizophrenia using family and population-based controls. Nat. Genet. *28*, 126-128.
- Skol,A.D., Young,K.A., Tsuang,D.W., Faraone,S.V., Haverstock,S.L., Bingham,S., Prabhudesai,S., Mena,F., Menon,A.S., Yu,C.E., Rundell,P., Pepple,J., Sauter,F., Baldwin,C., Weiss,D., Collins,J., Keith,T., Boehnke,M., Schellenberg,G.D., and Tsuang,M.T. (2003). Modest evidence for linkage and possible confirmation of association between NOTCH4 and schizophrenia in a large Veterans Affairs Cooperative Study sample. Am. J. Med. Genet. B Neuropsychiatr. Genet. *118*, 8-15.
- Smale,S.T. and Kadonaga,J.T. (2003). The RNA polymerase II core promoter. Annu. Rev. Biochem. *72*, 449-479.

- Smeraldi,E., Zanardi,R., Benedetti,F., Di Bella,D., Perez,J., and Catalano,M. (1998). Polymorphism within the promoter of the serotonin transporter gene and antidepressant efficacy of fluvoxamine. Mol. Psychiatry *3*, 508-511.
- Smigielski,E.M., Sirotkin,K., Ward,M., and Sherry,S.T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. *28*, 352-355.
- Smit,A.F. (1996). The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. *6*, 743-748.
- Sokal,R.R. and Rohlf,F.J. (1997). Biometry. (New York: W. H. Freeman and Co.).
- Sokoloff,P., Giros,B., Martres,M.P., Bouthenet,M.L., and Schwartz,J.C. (1990). Molecular cloning and characterization of a novel dopamine receptor (D3) as a target for neuroleptics. Nature *347*, 146-151.
- Sollars,V., Lu,X., Xiao,L., Wang,X., Garfinkel,M.D., and Ruden,D.M. (2003). Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. Nat. Genet. *33*, 70-74.
- Souery,D., Lipp,O., Mahieu,B., Mendelbaum,K., De,M., V, Van Broeckhoven,C., and Mendlewicz,J. (1996). Association study of bipolar disorder with candidate genes involved in catecholamine neurotransmission: DRD2, DRD3, DAT1, and TH genes. Am. J. Med. Genet. *67*, 551-555.
- Souery,D., Lipp,O., Rivelli,S.K., Massat,I., Serretti,A., Cavallini,C., Ackenheil,M., Adolfsson,R., Aschauer,H., Blackwood,D., Dam,H., Dikeos,D., Fuchshuber,S., Heiden,M., Jakovljevic,M., Kaneva,R., Kessing,L., Lerer,B., Lonnqvist,J., Mellerup,T., Milanova,V., Muir,W., Nylander,P.O., Oruc,L., and Mendlewicz,J. (1999). Tyrosine hydroxylase polymorphism and phenotypic heterogeneity in bipolar affective disorder: a multicenter association study. Am. J. Med. Genet. *88*, 527-532.
- Souery,D., Van Gestel,S., Massat,I., Blairy,S., Adolfsson,R., Blackwood,D., Del Favero,J., Dikeos,D., Jakovljevic,M., Kaneva,R., Lattuada,E., Lerer,B., Lilli,R., Milanova,V., Muir,W., Nothen,M., Oruc,L., Papadimitriou,G., Propping,P., Schulze,T., Serretti,A., Shapira,B., Smeraldi,E., Stefanis,C., Thomson,M., Van Broeckhoven,C., and Mendlewicz,J. (2001). Tryptophan hydroxylase polymorphism and suicidality in unipolar and bipolar affective disorders: a multicenter association study. Biol. Psychiatry *49*, 405-409.
- Spanagel,R. and Weiss,F. (1999). The dopamine hypothesis of reward: past and current status. Trends Neurosci. *22*, 521-527.
- Speek,M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell Biol. *21*, 1973-1985.
- Spellman,P.T. and Rubin,G.M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. *1*, 5.
- Spooren,W., Ballard,T., Gasparini,F., Amalric,M., Mutel,V., and Schreiber,R. (2003). Insight into the function of Group I and Group II metabotropic glutamate (mGlu) receptors: behavioural characterization and implications for the treatment of CNS disorders. Behav. Pharmacol. *14*, 257-277.
- Spurlock,G., Williams,J., McGuffin,P., Aschauer,H.N., Lenzinger,E., Fuchs,K., Sieghart,W.C., Meszaros,K., Fathi,N., Laurent,C., Mallet,J., Macciardi,F., Pedrini,S., Gill,M., Hawi,Z., Gibson,S., Jazin,E.E., Yang,H.T., Adolfsson,R., Pato,C.N., Dourado,A.M., and Owen,M.J. (1998). European Multicentre Association Study of Schizophrenia: a study of the DRD2 Ser311Cys and DRD3 Ser9Gly polymorphisms. Am. J. Med. Genet. *81*, 24-28.
- Srinivasan,J., Otto,G.W., Kahlow,U., Geisler,R., and Sommer,R.J. (2004). AppaDB: an AcedB database for the nematode satellite organism Pristionchus pacificus. Nucleic Acids Res. *32*, D421-D422.
- St Clair,D., Blackwood,D., Muir,W., Carothers,A., Walker,M., Spowart,G., Gosden,C., and Evans,H.J. (1990). Association within a family of a balanced autosomal translocation with major mental illness. Lancet *336*, 13-16.
- Stackman,R.W., Hammond,R.S., Linardatos,E., Gerlach,A., Maylie,J., Adelman,J.P., and Tzounopoulos,T. (2002). Small conductance Ca2+-activated K+ channels modulate synaptic plasticity and memory encoding. J. Neurosci. *22*, 10163-10171.

- Stahl,S.M. and Wets,K.M. (1988). Recent advances in drug delivery technology for neurology. Clin. Neuropharmacol. *11*, 1-17.
- Stambolic,V., Ruel,L., and Woodgett,J.R. (1996). Lithium inhibits glycogen synthase kinase-3 activity and mimics wingless signalling in intact cells. Curr. Biol. *6*, 1664-1668.
- Stamm,S., Ben Ari,S., Rafalska,I., Tang,Y., Zhang,Z., Toiber,D., Thanaraj,T.A., and Soreq,H. (2005). Function of alternative splicing. Gene *344*, 1-20.
- Stefansson,H., Sarginson,J., Kong,A., Yates,P., Steinthorsdottir,V., Gudfinnsson,E., Gunnarsdottir,S., Walker,N., Petursson,H., Crombie,C., Ingason,A., Gulcher,J.R., Stefansson,K., and St Clair,D. (2003). Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. Am. J. Hum. Genet. *72*, 83-87.
- Stefansson,H., Sigurdsson,E., Steinthorsdottir,V., Bjornsdottir,S., Sigmundsson,T., Ghosh,S., Brynjolfsson,J., Gunnarsdottir,S., Ivarsson,O., Chou,T.T., Hjaltason,O., Birgisdottir,B., Jonsson,H., Gudnadottir,V.G., Gudmundsdottir,E., Bjornsson,A., Ingvarsson,B., Ingason,A., Sigfusson,S., Hardardottir,H., Harvey,R.P., Lai,D., Zhou,M., Brunner,D., Mutel,V., Gonzalo,A., Lemke,G., Sainz,J., Johannesson,G., Andresson,T., Gudbjartsson,D., Manolescu,A., Frigge,M.L., Gurney,M.E., Kong,A., Gulcher,J.R., Petursson,H., and Stefansson,K. (2002). Neuregulin 1 and susceptibility to schizophrenia. Am. J. Hum. Genet. *71*, 877-892.
- Stefansson,H., Steinthorsdottir,V., Thorgeirsson,T.E., Gulcher,J.R., and Stefansson,K. (2004). Neuregulin 1 and schizophrenia. Ann. Med. *36*, 62-71.
- Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J., and Spieth,J. (2001). WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Res. *29*, 82-86.
- Stein,L.D. and Thierry-Mieg,J. (1998). Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. Genome Res. *8*, 1308-1315.
- Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M., and Cooper,D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. 2003. Jun. ;21(6.):577. -81. *21*, 577-581.
- Stephens,J.C., Schneider,J.A., Tanguay,D.A., Choi,J., Acharya,T., Stanley,S.E., Jiang,R., Messer,C.J., Chew,A., Han,J.H., Duan,J., Carr,J.L., Lee,M.S., Koshy,B., Kumar,A.M., Zhang,G., Newell,W.R., Windemuth,A., Xu,C., Kalbfleisch,T.S., Shaner,S.L., Arnold,K., Schulz,V., Drysdale,C.M., Nandabalan,K., Judson,R.S., Ruano,G., and Vovis,G.F. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science *293*, 489-493.
- Stine,O.C., Xu,J., Koskela,R., McMahon,F.J., Gschwend,M., Friddle,C., Clark,C.D., McInnis,M.G., Simpson,S.G., and Breschel,T.S. (1995). Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. Am. J. Hum. Genet. *57*, 1384-1394.
- Stober,G., Franzek,E., Lesch,K.P., and Beckmann,H. (1995). Periodic catatonia: a schizophrenic subtype with major gene effect and anticipation. Eur. Arch. Psychiatry Clin. Neurosci. *245*, 135-141.
- Stober,G., Haubitz,I., Franzek,E., and Beckmann,H. (1998a). Parent-of-origin effect and evidence for differential transmission in periodic catatonia. Psychiatr. Genet. *8*, 213-219.
- Stober,G., Jatzke,S., Meyer,J., Okladnova,O., Knapp,M., Beckmann,H., and Lesch,K.P. (1998b). Short CAG repeats within the hSKCa3 gene associated with schizophrenia: results of a family-based study. Neuroreport *9*, 3595-3599.
- Stober,G., Meyer,J., Nanda,I., Wienker,T.F., Saar,K., Jatzke,S., Schmid,M., Lesch,K.P., and Beckmann,H. (2000). hKCNN3 which maps to chromosome 1q21 is not the causative gene in periodic catatonia, a familial subtype of schizophrenia. Eur. Arch. Psychiatry Clin. Neurosci. *250*, 163-168.
- Stone,J.R. and Wray,G.A. (2001). Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. *18*, 1764-1770.
- Stopkova,P., Saito,T., Fann,C.S., Papolos,D.F., Vevera,J., Paclt,I., Zukov,I., Stryjer,R., Strous,R.D., and Lachman,H.M. (2003). Polymorphism screening of PIP5K2A: a candidate gene for chromosome 10p-linked psychiatric disorders. Am. J. Med. Genet. B Neuropsychiatr. Genet. *123*, 50-58.

- Stopkova,P., Saito,T., Papolos,D.F., Vevera,J., Paclt,I., Zukov,I., Bersson,Y.B., Margolis,B.A., Strous,R.D., and Lachman,H.M. (2004a). Identification of PIK3C3 promoter variant associated with bipolar disorder and schizophrenia. Biol. Psychiatry *55*, 981-988.
- Stopkova,P., Vevera,J., Paclt,I., Zukov,I., and Lachman,H.M. (2004b). Analysis of SYNJ1, a candidate gene for 21q22 linked bipolar disorder: a replication study. Psychiatry Res. *127*, 157-161.
- Stormo,G.D. (2000a). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.
- Stormo,G.D. (2000b). Gene-finding approaches for eukaryotes. Genome Res. *10*, 394-397.
- Strathdee,G., Sim,A., and Brown,R. (2004). Control of gene expression by CpG island methylation in normal cells. Biochem. Soc. Trans. *32*, 913-915.
- Straub,R.E., Jiang,Y., MacLean,C.J., Ma,Y., Webb,B.T., Myakishev,M.V., Harris-Kerr,C., Wormley,B., Sadek,H., Kadambi,B., Cesare,A.J., Gibberman,A., Wang,X., O'Neill,F.A., Walsh,D., and Kendler,K.S. (2002). Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. Am. J. Hum. Genet. *71*, 337-348.
- Straub,R.E., MacLean,C.J., O'Neill,F.A., Burke,J., Murphy,B., Duke,F., Shinkwin,R., Webb,B.T., Zhang,J., and Walsh,D. (1995). A potential vulnerability locus for schizophrenia on chromosome 6p24-22: evidence for genetic heterogeneity. Nat. Genet. *11*, 287-293.
- Strittmatter,W.J. and Roses,A.D. (1996). Apolipoprotein E and Alzheimer's disease. Annu. Rev. Neurosci. *19*, 53-77.
- Sultana,R., Yu,C.E., Yu,J., Munson,J., Chen,D., Hua,W., Estes,A., Cortes,F., de la,B.F., Yu,D., Haider,S.T., Trask,B.J., Green,E.D., Raskind,W.H., Disteche,C.M., Wijsman,E., Dawson,G., Storm,D.R., Schellenberg,G.D., and Villacres,E.C. (2002). Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. Genomics *80*, 129-134.
- Sumiyama,K., Kim,C.B., and Ruddle,F.H. (2001). An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics *71*, 260-262.
- Sundquist,K., Frank,G., and Sundquist,J. (2004). Urbanisation and incidence of psychosis and depression: follow-up study of 4.4 million women and men in Sweden. Br. J. Psychiatry *184*, 293-298.
- Susser,E., Neugebauer,R., Hoek,H.W., Brown,A.S., Lin,S., Labovitz,D., and Gorman,J.M. (1996). Schizophrenia after prenatal famine. Further evidence. Arch. Gen. Psychiatry *53*, 25-31.
- Suzuki,A., Yamada,R., Chang,X., Tokuhiro,S., Sawada,T., Suzuki,M., Nagasaki,M., Nakayama-Hamada,M., Kawaida,R., Ono,M., Ohtsuki,M., Furukawa,H., Yoshino,S., Yukioka,M., Tohma,S., Matsubara,T., Wakitani,S., Teshima,R., Nishioka,Y., Sekine,A., Iida,A., Takahashi,A., Tsunoda,T., Nakamura,Y., and Yamamoto,K. (2003). Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat. Genet. *34*, 395-402.
- Svejstrup,J.Q. (2004). The RNA polymerase II transcription cycle: cycling through chromatin. Biochim. Biophys. Acta *1677*, 64-73.
- Svensson,T.H. (2000). Dysfunctional brain dopamine systems induced by psychotomimetic NMDA-receptor antagonists and the effects of antipsychotic drugs. Brain Res. Brain Res. Rev. *31*, 320-329.
- Swift,M. and Swift,R.G. (2000). Psychiatric disorders and mutations at the Wolfram syndrome locus. Biol. Psychiatry *47*, 787-793.
- Swift-Scanlan,T., Coughlin,J.M., Lan,T.H., Potash,J.B., Ingersoll,R.G., Depaulo,R., Jr., Ross,C.A., and McInnis,M.G. (2005). Characterization of CTG/CAG repeats on chromosome 18: a study of bipolar disorder. Psychiatr. Genet. *15*, 91-99.
- Syagailo,Y.V., Stober,G., Grassle,M., Reimer,E., Knapp,M., Jungkunz,G., Okladnova,O., Meyer,J., and Lesch,K.P. (2001). Association analysis of the functional monoamine oxidase A gene promoter polymorphism in psychiatric disorders. Am. J. Med. Genet. *105*, 168-171.

- Tachikawa,H., Harada,S., Kawanishi,Y., Okubo,T., and Shiraishi,H. (2000). Novel polymorphisms of the human cholecystokinin A receptor gene: an association analysis with schizophrenia. Am. J. Med. Genet. *96*, 141-145.
- Tachikawa,H., Harada,S., Kawanishi,Y., Okubo,T., and Suzuki,T. (2001). Linked polymorphisms (-333G>T and -286A>G) in the promoter region of the CCK-A receptor gene may be associated with schizophrenia. Psychiatry Res. *103*, 147-155.
- Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L., and Jones,R.T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. *203*, 439-455.
- Tahir,E., Yazgan,Y., Cirakoglu,B., Ozbay,F., Waldman,I., and Asherson,P.J. (2000). Association and linkage of DRD4 and DRD5 with attention deficit hyperactivity disorder (ADHD) in a sample of Turkish children. Mol. Psychiatry *5*, 396-404.
- Taillon-Miller,P., Gu,Z., Li,Q., Hillier,L., and Kwok,P.Y. (1998). Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res. *8*, 748-754.
- Taipale,M., Kaminen,N., Nopola-Hemmi,J., Haltia,T., Myllyluoma,B., Lyytinen,H., Muller,K., Kaaranen,M., Lindsberg,P.J., Hannula-Jouppi,K., and Kere,J. (2003). A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. Proc. Natl. Acad. Sci. U. S. A. *100*, 11553-11558.
- Takahashi,S., Cui,Y.H., Kojima,T., Han,Y.H., Yu,S.Y., Tanabe,E., Yara,K., Matsuura,M., Matsushima,E., Nakayama,J., Arinami,T., Shen,Y.C., Faraone,S.V., and Tsuang,M.T. (2003b). Family-based association study of the NOTCH4 gene in schizophrenia using Japanese and Chinese samples. Biol. Psychiatry *54*, 129-135.
- Takahashi,S., Cui,Y.H., Kojima,T., Han,Y.H., Yu,S.Y., Tanabe,E., Yara,K., Matsuura,M., Matsushima,E., Nakayama,J., Arinami,T., Shen,Y.C., Faraone,S.V., and Tsuang,M.T. (2003a). Family-based association study of the NOTCH4 gene in schizophrenia using Japanese and Chinese samples. Biol. Psychiatry *54*, 129-135.
- Takai-Igarashi,T., Nadaoka,Y., and Kaminuma,T. (1998). A database for cell signaling networks. J. Comput. Biol. *5*, 747-754.
- Tang,A.H., Franklin,S.R., Himes,C.S., Smith,M.W., and Tenbrink,R.E. (1997). PNU-96415E, a potential antipsychotic agent with clozapine-like pharmacological properties. J. Pharmacol. Exp. Ther. *281*, 440-447.
- Tang,J.X., Chen,W.Y., He,G., Zhou,J., Gu,N.F., Feng,G.Y., and He,L. (2004). Polymorphisms within 5' end of the Neuregulin 1 gene are genetically associated with schizophrenia in the Chinese population. Mol. Psychiatry *9*, 11-12.
- Tang,J.X., Zhou,J., Fan,J.B., Li,X.W., Shi,Y.Y., Gu,N.F., Feng,G.Y., Xing,Y.L., Shi,J.G., and He,L. (2003). Family-based association study of DTNBP1 in 6p22.3 and schizophrenia. Mol. Psychiatry *8*, 717-718.
- Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281-285.
- Taylor,M.S., Devon,R.S., Millar,J.K., and Porteous,D.J. (2003). Evolutionary constraints on the Disrupted in Schizophrenia locus. Genomics *81*, 67-77.
- Teisberg,P. (1995). The genetic background of anticipation. J. R. Soc. Med. *88*, 185-187.
- The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell *72*, 971-983.
- The International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789-796.
- Theuns,J., Remacle,J., Killick,R., Corsmit,E., Vennekens,K., Huylebroeck,D., Cruts,M., and Van Broeckhoven,C. (2003). Alzheimer-associated C allele of the promoter polymorphism -22C>T causes a critical neuron-specific decrease of presenilin 1 expression. Hum. Mol. Genet. *12*, 869-877.
- Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C., Maskeri,B.,

Hansen,N.F., Schwartz,M.S., Weber,R.J., Kent,W.J., Karolchik,D., Bruen,T.C., Bevan,R., Cutler,D.J., Schwartz,S., Elnitski,L., Idol,J.R., Prasad,A.B., Lee-Lin,S.Q., Maduro,V.V., Summers,T.J., Portnoy,M.E., Dietrich,N.L., Akhter,N., Ayele,K., Benjamin,B., Cariaga,K., Brinkley,C.P., Brooks,S.Y., Granite,S., Guan,X., Gupta,J., Haghighi,P., Ho,S.L., Huang,M.C., Karlins,E., Laric,P.L., Legaspi,R., Lim,M.J., Maduro,Q.L., Masiello,C.A., Mastrian,S.D., McCloskey,J.C., Pearson,R., Stantripop,S., Tiongson,E.E., Tran,J.T., Tsurgeon,C., Vogt,J.L., Walker,M.A., Wetherby,K.D., Wiggins,L.S., Young,A.C., Zhang,L.H., Osoegawa,K., Zhu,B., Zhao,B., Shu,C.L., De Jong,P.J., Lawrence,C.E., Smit,A.F., Chakravarti,A., Haussler,D., Green,P., Miller,W., and Green,E.D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature *424*, 788-793.

- Thompson,J.D., Higgins,D.G., and Gibson,T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673-4680.
- Thomson,P.A., Wray,N.R., Millar,J.K., Evans,K.L., Hellard,S.L., Condie,A., Muir,W.J., Blackwood,D.H., and Porteous,D.J. (2005). Association between the TRAX/DISC locus and both bipolar disorder and schizophrenia in the Scottish population. Mol. Psychiatry *10*, 657-668.
- Ti,H. and Suwen,N. (1966). The yellow Emperor's classic of internal medicine. (Berkeley: University of California Press).
- Ting,C.N., Rosenberg,M.P., Snow,C.M., Samuelson,L.C., and Meisler,M.H. (1992). Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. *6*, 1457-1465.
- Tiret,L., Poirier,O., Nicaud,V., Barbaux,S., Herrmann,S.M., Perret,C., Raoux,S., Francomme,C., Lebard,G., Tregouet,D., and Cambien,F. (2002). Heterogeneity of linkage disequilibrium in human genes has implications for association studies of common diseases. Hum. Mol. Genet. *11*, 419-429.
- Tochigi,M., Zhang,X., Umekage,T., Ohashi,J., Kato,C., Marui,T., Otowa,T., Hibino,H., Otani,T., Kohda,K., Liu,S., Kato,N., Tokunaga,K., and Sasaki,T. (2004). Association of six polymorphisms of the NOTCH4 gene with schizophrenia in the Japanese population. Am. J. Med. Genet. B Neuropsychiatr. Genet. *128*, 37-40.
- Tokuhiro,S., Yamada,R., Chang,X., Suzuki,A., Kochi,Y., Sawada,T., Suzuki,M., Nagasaki,M., Ohtsuki,M., Ono,M., Furukawa,H., Nagashima,M., Yoshino,S., Mabuchi,A., Sekine,A., Saito,S., Takahashi,A., Tsunoda,T., Nakamura,Y., and Yamamoto,K. (2003). An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nat. Genet. *35*, 341-348.
- Toma,D.P., White,K.P., Hirsch,J., and Greenspan,R.J. (2002). Identification of genes involved in Drosophila melanogaster geotaxis, a complex behavioral trait. Nat. Genet. *31*, 349-353.
- Torrey,E.F., Miller,J., Rawlings,R., and Yolken,R.H. (1997). Seasonality of births in schizophrenia and bipolar disorder: a review of the literature. Schizophr. Res. *28*, 1-38.
- Townsend,J.P., Cavalieri,D., and Hartl,D.L. (2003). Population genetic variation in genome-wide gene expression. Mol. Biol. Evol. *20*, 955-963.
- Tsai,S.J., OuYang,W.C., and Hong,C.J. (2002). Association for serotonin transporter gene variable number tandem repeat polymorphism and schizophrenic disorders. Neuropsychobiology *45*, 131-133.
- Tsui-Pierchala,B.A., Encinas,M., Milbrandt,J., and Johnson,E.M., Jr. (2002). Lipid rafts in neuronal signaling and function. Trends Neurosci. *25*, 412-417.
- Tsutsumi,T., Holmes,S.E., McInnis,M.G., Sawa,A., Callahan,C., DePaulo,J.R., Ross,C.A., DeLisi,L.E., and Margolis,R.L. (2004). Novel CAG/CTG repeat expansion mutations do not contribute to the genetic risk for most cases of bipolar disorder or schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *124*, 15-19.
- Ueda,H., Howson,J.M., Esposito,L., Heward,J., Snook,H., Chamberlain,G., Rainbow,D.B., Hunter,K.M., Smith,A.N., Di Genova,G., Herr,M.H., Dahlman,I., Payne,F., Smyth,D.,

Lowe,C., Twells,R.C., Howlett,S., Healy,B., Nutland,S., Rance,H.E., Everett,V., Smink,L.J., Lam,A.C., Cordell,H.J., Walker,N.M., Bordin,C., Hulme,J., Motzo,C., Cucca,F., Hess,J.F., Metzker,M.L., Rogers,J., Gregory,S., Allahabadia,A., Nithiyananthan,R., Tuomilehto-Wolf,E., Tuomilehto,J., Bingley,P., Gillespie,K.M., Undlien,D.E., Ronningen,K.S., Guja,C., Ionescu-Tirgoviste,C., Savage,D.A., Maxwell,A.P., Carson,D.J., Patterson,C.C., Franklyn,J.A., Clayton,D.G., Peterson,L.B., Wicker,L.S., Todd,J.A., and Gough,S.C. (2003). Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature *423*, 506-511.

- Ujike,H., Yamamoto,A., Tanaka,Y., Takehisa,Y., Takaki,M., Taked,T., Kodama,M., and Kuroda,S. (2001). Association study of CAG repeats in the KCNN3 gene in Japanese patients with schizophrenia, schizoaffective disorder and bipolar disorder. Psychiatry Res. *101*, 203-207.
- Usiskin,S.I., Nicolson,R., Krasnewich,D.M., Yan,W., Lenane,M., Wudarsky,M., Hamburger,S.D., and Rapoport,J.L. (1999). Velocardiofacial syndrome in childhood-onset schizophrenia. J. Am. Acad. Child Adolesc. Psychiatry *38*, 1536-1543.
- Ustun,T.B. (1999). The global burden of mental disorders. Am. J. Public Health *89*, 1315-1318.
- Uyttendaele,H., Ho,J., Rossant,J., and Kitajewski,J. (2001). Vascular patterning defects associated with expression of activated Notch4 in embryonic endothelium. Proc. Natl. Acad. Sci. U. S. A. *98*, 5643-5648.
- Uyttendaele,H., Marazzi,G., Wu,G., Yan,Q., Sassoon,D., and Kitajewski,J. (1996). Notch4/int-3, a mammary proto-oncogene, is an endothelial cell-specific mammalian Notch gene. Development *122*, 2251-2259.
- Van Den Bogaert,A., Schumacher, J., Schulze,T.G., Otte,A.C., Ohlraun,S., Kovalenko,S., Becker,T., Freudenberg,J., Jonsson,E.G., Mattila-Evenden,M., Sedvall,G.C., Czerski,P.M., Kapelski,P., Hauser,J., Maier,W., Rietschel,M., Propping,P., Nothen,M.M., Cichon,S. (2003). The DTNBP1 (dysbindin) gene contributes to schizophrenia, depending on family history of the disease. Am. J. Hum. Genet. *73*:1438-43
- van den Oord,E.J., Sullivan,P.F., Jiang,Y., Walsh,D., O'Neill,F.A., Kendler,K.S., and Riley,B.P. (2003). Identification of a high-risk haplotype for the dystrobrevin binding protein 1 (DTNBP1) gene in the Irish study of high-density schizophrenia families. Mol. Psychiatry *8*, 499-510.
- van Helden,J., Andre,B., and Collado-Vides,J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. *281*, 827-842.
- van Helden,J., Rios,A.F., and Collado-Vides,J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. *28*, 1808-1818.
- van Leeuwen,F.W., de Kleijn,D.P., van den Hurk,H.H., Neubauer,A., Sonnemans,M.A., Sluijs,J.A., Koycu,S., Ramdjielal,R.D., Salehi,A., Martens,G.J., Grosveld,F.G., Peter,J., Burbach,H., and Hol,E.M. (1998). Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. Science *279*, 242-247.
- van Overveld,P.G., Lemmers,R.J., Sandkuijl,L.A., Enthoven,L., Winokur,S.T., Bakels,F., Padberg,G.W., van Ommen,G.J., Frants,R.R., and van der Maarel,S.M. (2003). Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. Nat. Genet. *35*, 315-317.
- van Rijn,S., Aleman,A., Swaab,H., and Kahn,R.S. (2005). Neurobiology of emotion and high risk for schizophrenia: role of the amygdala and the X-chromosome. Neurosci. Biobehav. Rev. *29*, 385-397.
- Van Tol,H.H., Bunzow,J.R., Guan,H.C., Sunahara,R.K., Seeman,P., Niznik,H.B., and Civelli,O. (1991). Cloning of the gene for a human dopamine D4 receptor with high affinity for the antipsychotic clozapine. Nature *350*, 610-614.
- Van Tol,H.H., Wu,C.M., Guan,H.C., Ohara,K., Bunzow,J.R., Civelli,O., Kennedy,J., Seeman,P., Niznik,H.B., and Jovanovic,V. (1992). Multiple dopamine D4 receptor variants in the human population. Nature *358*, 149-152.

- Vanakoski,J., Virkkunen,M., Naukkarinen,H., and Goldman,D. (2001). No association of CCK and CCK(B) receptor polymorphisms with alcohol dependence. Psychiatry Res. *102*, 1-7.
- Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., Amanatides,P., Ballew,R.M., Huson,D.H., Wortman,J.R., Zhang,Q., Kodira,C.D., Zheng,X.H., Chen,L., Skupski,M., Subramanian,G., Thomas,P.D., Zhang,J., Gabor Miklos,G.L., Nelson,C., Broder,S., Clark,A.G., Nadeau,J., McKusick,V.A., Zinder,N., Levine,A.J., Roberts,R.J., Simon,M., Slayman,C., Hunkapiller,M., Bolanos,R., Delcher,A., Dew,I., Fasulo,D., Flanigan,M., Florea,L., Halpern,A., Hannenhalli,S., Kravitz,S., Levy,S., Mobarry,C., Reinert,K., Remington,K., Abu-Threideh,J., Beasley,E., Biddick,K., Bonazzi,V., Brandon,R., Cargill,M., Chandramouliswaran,I., Charlab,R., Chaturvedi,K., Deng,Z., Di,F., V, Dunn,P., Eilbeck,K., Evangelista,C., Gabrielian,A.E., Gan,W., Ge,W., Gong,F., Gu,Z., Guan,P., Heiman,T.J., Higgins,M.E., Ji,R.R., Ke,Z., Ketchum,K.A., Lai,Z., Lei,Y., Li,Z., Li,J., Liang,Y., Lin,X., Lu,F., Merkulov,G.V., Milshina,N., Moore,H.M., Naik,A.K., Narayan,V.A., Neelam,B., Nusskern,D., Rusch,D.B., Salzberg,S., Shao,W., Shue,B., Sun,J., Wang,Z., Wang,A., Wang,X., Wang,J., Wei,M., Wides,R., Xiao,C., Yan,C., Yao,A., Ye,J., Zhan,M., Zhang,W., Zhang,H., Zhao,Q., Zheng,L., Zhong,F., Zhong,W., Zhu,S., Zhao,S., Gilbert,D., Baumhueter,S., Spier,G., Carter,C., Cravchik,A., Woodage,T., Ali,F., An,H., Awe,A., Baldwin,D., Baden,H., Barnstead,M., Barrow,I., Beeson,K., Busam,D., Carver,A., Center,A., Cheng,M.L., Curry,L., Danaher,S., Davenport,L., Desilets,R., Dietz,S., Dodson,K., Doup,L., Ferriera,S., Garg,N., Gluecksmann,A., Hart,B., Haynes,J., Haynes,C., Heiner,C., Hladun,S., Hostin,D., Houck,J., Howland,T., Ibegwam,C., Johnson,J., Kalush,F., Kline,L., Koduru,S., Love,A., Mann,F., May,D., McCawley,S., McIntosh,T., McMullen,I., Moy,M., Moy,L., Murphy,B., Nelson,K., Pfannkoch,C., Pratts,E., Puri,V., Qureshi,H., Reardon,M., Rodriguez,R., Rogers,Y.H., Romblad,D., Ruhfel,B., Scott,R., Sitter,C., Smallwood,M., Stewart,E., Strong,R., Suh,E., Thomas,R., Tint,N.N., Tse,S., Vech,C., Wang,G., Wetter,J., Williams,S., Williams,M., Windsor,S., Winn-Deen,E., Wolfe,K., Zaveri,J., Zaveri,K., Abril,J.F., Guigo,R., Campbell,M.J., Sjolander,K.V., Karlak,B., Kejariwal,A., Mi,H., Lazareva,B., Hatton,T., Narechania,A., Diemer,K., Muruganujan,A., Guo,N., Sato,S., Bafna,V., Istrail,S., Lippert,R., Schwartz,R., Walenz,B., Yooseph,S., Allen,D., Basu,A., Baxendale,J., Blick,L., Caminha,M., Carnes-Stine,J., Caulk,P., Chiang,Y.H., Coyne,M., Dahlke,C., Mays,A., Dombroski,M., Donnelly,M., Ely,D., Esparham,S., Fosler,C., Gire,H., Glanowski,S., Glasser,K., Glodek,A., Gorokhov,M., Graham,K., Gropman,B., Harris,M., Heil,J., Henderson,S., Hoover,J., Jennings,D., Jordan,C., Jordan,J., Kasha,J., Kagan,L., Kraft,C., Levitsky,A., Lewis,M., Liu,X., Lopez,J., Ma,D., Majoros,W., McDaniel,J., Murphy,S., Newman,M., Nguyen,T., Nguyen,N., and Nodell,M. (2001). The sequence of the human genome. Science *291*, 1304-1351.
- Veraksa,A., Del Campo,M., and McGinnis,W. (2000). Developmental patterning genes and their conserved functions: from model organisms to humans. Mol. Genet. Metab. *69*, 85-100.
- Verschure,P.J. (2004). Positioning the genome within the nucleus. Biol. Cell *96*, 569-577.
- Vincent,J.B., Herbrick,J.A., Gurling,H.M., Bolton,P.F., Roberts,W., and Scherer,S.W. (2000a). Identification of a novel gene on chromosome 7q31 that is interrupted by a translocation breakpoint in an autistic individual. Am. J. Hum. Genet. *67*, 510-514.
- Vincent,J.B., Paterson,A.D., Strong,E., Petronis,A., and Kennedy,J.L. (2000b). The unstable trinucleotide repeat story of major psychosis. Am. J. Med. Genet. *97*, 77-97.
- Vincent,J.B., Petek,E., Thevarkunnel,S., Kolozsvari,D., Cheung,J., Patel,M., and Scherer,S.W. (2002). The RAY1/ST7 tumor-suppressor locus on chromosome 7q31 represents a complex multi-transcript system. Genomics *80*, 283-294.
- Visscher,P.M., Haley,C.S., Heath,S.C., Muir,W.J., and Blackwood,D.H. (1999). Detecting QTLs for uni- and bipolar disorder using a variance component method. Psychiatr. Genet. *9*, 75-84.
- Visscher,P.M., Yazdi,M.H., Jackson,A.D., Schalling,M., Lindblad,K., Yuan,Q.P., Porteous,D., Muir,W.J., and Blackwood,D.H. (2001). Genetic survival analysis of age-at-

onset of bipolar disorder: evidence for anticipation or cohort effect in families. Psychiatr. Genet. *11*, 129-137.

- Vogel,M., Pfeifer,S., Schaub,R.T., Grabe,H.J., Barnow,S., Freyberger,H.J., and Cascorbi,I. (2004). Decreased levels of dopamine D3 receptor mRNA in schizophrenic and bipolar patients. Neuropsychobiology *50*, 305-310.
- Wall,L.T., Christiansen,T., and Schwartz,R. (2000). Programming Perl. (Sebastopol, CA: O'Reilly).
- Walsh,S., Anderson,M., and Cartinhour,S.W. (1998). ACEDB: a database for genome information. Methods Biochem. Anal. *39*, 299-318.
- Walters,J.R., Bergstrom,D.A., Carlson,J.H., Chase,T.N., and Braun,A.R. (1987). D1 dopamine receptor activation required for postsynaptic expression of D2 agonist effects. Science *236*, 719-722.
- Wang,J., Si,Y.M., Liu,Z.L., and Yu,L. (2003). Cholecystokinin, cholecystokinin-A receptor and cholecystokinin-B receptor gene polymorphisms in Parkinson's disease. Pharmacogenetics *13*, 365-369.
- Wang,S., Sun,C.E., Walczak,C.A., Ziegle,J.S., Kipps,B.R., Goldin,L.R., and Diehl,S.R. (1995). Evidence for a susceptibility locus for schizophrenia on chromosome 6pter-p22. Nat. Genet. *10*, 41-46.
- Wang,W.C., Anand,S., Powell,D.R., Pawashe,A.B., Amemiya,C.T., and Shashikant,C.S. (2004). Comparative cis-regulatory analyses identify new elements of the mouse Hoxc8 early enhancer. J. Exp. Zoolog. B Mol. Dev. Evol. *302*, 436-445.
- Wang,Z., Valdes,J., Noyes,R., Zoega,T., and Crowe,R.R. (1998). Possible association of a cholecystokinin promotor polymorphism (CCK-36CT) with panic disorder. Am. J. Med. Genet. *81*, 228-234.
- Wang,Z., Wassink,T., Andreasen,N.C., and Crowe,R.R. (2002). Possible association of a cholecystokinin promoter variant to schizophrenia. Am. J. Med. Genet. *114*, 479-482.
- Wank,S.A., Harkins,R., Jensen,R.T., Shapira,H., De Weerth,A., and Slattery,T. (1992). Purification, molecular cloning, and functional expression of the cholecystokinin receptor from rat pancreas. Proc. Natl. Acad. Sci. U. S. A *89*, 3125-3129.
- Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W., and Lawrence,C.E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. *26*, 225-228.
- Wassink,T.H., Nopoulos,P., Pietila,J., Crowe,R.R., and Andreasen,N.C. (2003). NOTCH4 and the frontal lobe in schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. *118*, 1-7.
- Waterston,R. and Sulston,J. (1995). The genome of Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A *92*, 10836-10840.
- Waterston,R. and Sulston,J.E. (1998). The Human Genome Project: reaching the finish line. Science *282*, 53-54.
- Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., Antonarakis,S.E., Attwood,J., Baertsch,R., Bailey,J., Barlow,K., Beck,S., Berry,E., Birren,B., Bloom,T., Bork,P., Botcherby,M., Bray,N., Brent,M.R., Brown,D.G., Brown,S.D., Bult,C., Burton,J., Butler,J., Campbell,R.D., Carninci,P., Cawley,S., Chiaromonte,F., Chinwalla,A.T., Church,D.M., Clamp,M., Clee,C., Collins,F.S., Cook,L.L., Copley,R.R., Coulson,A., Couronne,O., Cuff,J., Curwen,V., Cutts,T., Daly,M., David,R., Davies,J., Delehaunty,K.D., Deri,J., Dermitzakis,E.T., Dewey,C., Dickens,N.J., Diekhans,M., Dodge,S., Dubchak,I., Dunn,D.M., Eddy,S.R., Elnitski,L., Emes,R.D., Eswara,P., Eyras,E., Felsenfeld,A., Fewell,G.A., Flicek,P., Foley,K., Frankel,W.N., Fulton,L.A., Fulton,R.S., Furey,T.S., Gage,D., Gibbs,R.A., Glusman,G., Gnerre,S., Goldman,N., Goodstadt,L., Grafham,D., Graves,T.A., Green,E.D., Gregory,S., Guigo,R., Guyer,M., Hardison,R.C., Haussler,D., Hayashizaki,Y., Hillier,L.W., Hinrichs,A., Hlavina,W., Holzer,T., Hsu,F., Hua,A., Hubbard,T., Hunt,A., Jackson,I., Jaffe,D.B., Johnson,L.S., Jones,M., Jones,T.A., Joy,A., Kamal,M., Karlsson,E.K., Karolchik,D., Kasprzyk,A., Kawai,J., Keibler,E., Kells,C., Kent,W.J., Kirby,A., Kolbe,D.L., Korf,I., Kucherlapati,R.S., Kulbokas,E.J., Kulp,D., Landers,T., Leger,J.P., Leonard,S., Letunic,I., Levine,R., Li,J., Li,M., Lloyd,C., Lucas,S., Ma,B., Maglott,D.R., Mardis,E.R., Matthews,L.,

Mauceli,E., Mayer,J.H., McCarthy,M., McCombie,W.R., McLaren,S., McLay,K., McPherson,J.D., Meldrim,J., Meredith,B., Mesirov,J.P., Miller,W., Miner,T.L., Mongin,E., Montgomery,K.T., Morgan,M., Mott,R., Mullikin,J.C., Muzny,D.M., Nash,W.E., Nelson,J.O., Nhan,M.N., Nicol,R., Ning,Z., Nusbaum,C., O'Connor,M.J., Okazaki,Y., Oliver,K., Overton-Larty,E., Pachter,L., Parra,G., Pepin,K.H., Peterson,J., Pevzner,P., Plumb,R., Pohl,C.S., Poliakov,A., Ponce,T.C., Ponting,C.P., Potter,S., Quail,M., Reymond,A., Roe,B.A., Roskin,K.M., Rubin,E.M., Rust,A.G., Santos,R., Sapojnikov,V., Schultz,B., Schultz,J., Schwartz,M.S., Schwartz,S., Scott,C., Seaman,S., Searle,S., Sharpe,T., Sheridan,A., Shownkeen,R., Sims,S., Singer,J.B., Slater,G., Smit,A., Smith,D.R., Spencer,B., Stabenau,A., Stange-Thomann,N., Sugnet,C., Suyama,M., Tesler,G., Thompson,J., Torrents,D., Trevaskis,E., Tromp,J., Ucla,C., Ureta-Vidal,A., Vinson,J.P., Von Niederhausern,A.C., Wade,C.M., Wall,M., Weber,R.J., Weiss,R.B., Wendl,M.C., West,A.P., Wetterstrand,K., Wheeler,R., Whelan,S., Wierzbowski,J., Willey,D., Williams,S., Wilson,R.K., Winter,E., Worley,K.C., Wyman,D., Yang,S., Yang,S.P., Zdobnov,E.M., Zody,M.C., and Lander,E.S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

- Watkins,W.S., Rogers,A.R., Ostler,C.T., Wooding,S., Bamshad,M.J., Brassington,A.M., Carroll,M.L., Nguyen,S.V., Walker,J.A., Prasad,B.V., Reddy,P.G., Das,P.K., Batzer,M.A., and Jorde,L.B. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res. *13*, 1607-1618.
- Waugh,M.G., Minogue,S., Anderson,J.S., dos,S.M., and Hsuan,J.J. (2001). Signalling and non-caveolar rafts. Biochem. Soc. Trans. *29*, 509-511.
- Weatherbee,S.D., Nijhout,H.F., Grunert,L.W., Halder,G., Galant,R., Selegue,J., and Carroll,S. (1999). Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. Curr. Biol. *9*, 109-115.
- Wei,J. and Hemmings,G.P. (1999). The CCK-A receptor gene possibly associated with auditory hallucinations in schizophrenia. Eur. Psychiatry *14*, 67-70.
- Wei,J. and Hemmings,G.P. (2000). The NOTCH4 locus is associated with susceptibility to schizophrenia. Nat. Genet. *25*, 376-377.
- Wei,Y.J., Sun,H.Q., Yamamoto,M., Wlodarski,P., Kunii,K., Martinez,M., Barylko,B., Albanesi,J.P., and Yin,H.L. (2002). Type II phosphatidylinositol 4-kinase beta is a cytosolic and peripheral membrane protein that is recruited to the plasma membrane and activated by Rac-GTP. J. Biol. Chem. *277*, 46586-46593.
- Weickert,C.S., Straub,R.E., McClintock,B.W., Matsumoto,M., Hashimoto,R., Hyde,T.M., Herman,M.M., Weinberger,D.R., and Kleinman,J.E. (2004). Human dysbindin (DTNBP1) gene expression in normal brain and in schizophrenic prefrontal cortex and midbrain. Arch. Gen. Psychiatry *61*, 544-555.
- Weissman,M.M., Leaf,P.J., Tischler,G.L., Blazer,D.G., Karno,M., Bruce,M.L., and Florio,L.P. (1988). Affective disorders in five United States communities. Psychol. Med. *18*, 141-153.
- Weixel,K.M., Blumental-Perry,A., Watkins,S.C., Aridor,M., and Weisz,O.A. (2005). Distinct Golgi populations of phosphatidylinositol 4-phosphate regulated by phosphatidylinositol 4-kinases. J. Biol. Chem. *280*, 10501-10508.
- Wenk,M.R., Pellegrini,L., Klenchin,V.A., Di Paolo,G., Chang,S., Daniell,L., Arioka,M., Martin,T.F., and De Camilli,P. (2001). PIP kinase Igamma is the major PI(4,5)P(2) synthesizing enzyme at the synapse. Neuron *32*, 79-88.
- Werner,T. (2000). Identification and functional modelling of DNA sequence elements of transcription. Brief. Bioinform. *1*, 372-380.
- Werner,T. (2003a). Promoters can contribute to the elucidation of protein function. Trends Biotechnol. *21*, 9-13.
- Werner,T. (2003b). The state of the art of mammalian promoter recognition. Brief. Bioinform. *4*, 22-30.
- Werner,T., Fessele,S., Maier,H., and Nelson,P.J. (2003). Computer modeling of promoter organization as a tool to study transcriptional coregulation. FASEB J. *17*, 1228-1237.

- Werth,V.P., Zhang,W., Dortzbach,K., and Sullivan,K. (2000). Association of a promoter polymorphism of tumor necrosis factor-alpha with subacute cutaneous lupus erythematosus and distinct photoregulation of transcription. J. Invest Dermatol. *115*, 726-730.
- White R.J. (2001). Gene transcription: mechanisms and control. (Oxford: Blackwell Science).
- Whitney,A.R., Diehn,M., Popper,S.J., Alizadeh,A.A., Boldrick,J.C., Relman,D.A., and Brown,P.O. (2003). Individuality and variation in gene expression patterns in human blood. Proc. Natl. Acad. Sci. U. S. A. *100*, 1896-1901.
- Wible,C.G., Anderson,J., Shenton,M.E., Kricun,A., Hirayasu,Y., Tanaka,S., Levitt,J.J., O'Donnell,B.F., Kikinis,R., Jolesz,F.A., and McCarley,R.W. (2001). Prefrontal cortex, negative symptoms, and schizophrenia: an MRI study. Psychiatry Res. *108*, 65-78.
- Wiedemann,C., Schafer,T., Burger,M.M., and Sihra,T.S. (1998). An essential role for a small synaptic vesicle-associated phosphatidylinositol 4-kinase in neurotransmitter release. J. Neurosci. *18*, 5594-5602.
- Wildenauer,D.B., Schwab,S.G., Maier,W., and Detera-Wadleigh,S.D. (1999). Do schizophrenia and affective disorder share susceptibility genes? Schizophr. Res. *39*, 107-111.
- Williams,N.M., Cardno,A.G., Murphy,K.C., Jones,L.A., Asherson,P., McGuffin,P., and Owen,M.J. (1997). Association between schizophrenia and a microsatellite polymorphism at the dopamine D5 receptor gene. Psychiatr. Genet. *7*, 83-85.
- Williams,N.M., Rees,M.I., Holmans,P., Norton,N., Cardno,A.G., Jones,L.A., Murphy,K.C., Sanders,R.D., McCarthy,G., Gray,M.Y., Fenton,I., McGuffin,P., and Owen,M.J. (1999). A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs. Hum. Mol. Genet. *8*, 1729-1739.
- Williams,R.S., Cheng,L., Mudge,A.W., and Harwood,A.J. (2002). A common mechanism of action for three mood-stabilizing drugs. Nature *417*, 292-295.
- Williams,H.J., Williams,N., Spurlock,G., Norton,N., Ivanov,D., McCreadie,R.G., Preece,A., Sharkey,V., Jones,S., Zammit,S., Nikolov,I., Kehaiov,I., Thapar,A., Murphy,K.C., Kirov,G., Owen,M.J., and O'Donovan,M.C. (2003a). Association between PRODH and schizophrenia is not confirmed. Mol. Psychiatry *8*, 644-645.
- Williams,N.M., Norton,N., Williams,H., Ekholm,B., Hamshere,M.L., Lindblom,Y., Chowdari,K.V., Cardno,A.G., Zammit,S., Jones,L.A., Murphy,K.C., Sanders,R.D., McCarthy,G., Gray,M.Y., Jones,G., Holmans,P., Nimgaonkar,V., Adolfson,R., Osby,U., Terenius,L., Sedvall,G., O'Donovan,M.C., and Owen,M.J. (2003b). A systematic genomewide linkage study in 353 sib pairs with schizophrenia. Am. J. Hum. Genet. *73*, 1355-1367.
- Williams,N.M., Preece,A., Spurlock,G., Norton,N., Williams,H.J., Zammit,S., O'Donovan,M.C., and Owen,M.J. (2003c). Support for genetic variation in neuregulin 1 and susceptibility to schizophrenia. Mol. Psychiatry *8*, 485-487.
- Williams,N.M., Preece,A., Morris,D.W., Spurlock,G., Bray,N.J., Stephens,M., Norton,N., Williams,H., Clement,M., Dwyer,S., Curran,C., Wilkinson,J., Moskvina,V., Waddington,J.L., Gill,M., Corvin,A.P., Zammit,S., Kirov,G., Owen,M.J., and O'Donovan,M.C. (2004a). Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin binding protein gene (DTNBP1). Arch. Gen. Psychiatry *61*, 336-344.
- Williams,N.M., Preece,A., Spurlock,G., Norton,N., Williams,H.J., McCreadie,R.G., Buckland,P., Sharkey,V., Chowdari,K.V., Zammit,S., Nimgaonkar,V., Kirov,G., Owen,M.J., and O'Donovan,M.C. (2004b). Support for RGS4 as a susceptibility gene for schizophrenia. Biol. Psychiatry *55*, 192-195.
- Williams,R., Ryves,W.J., Dalton,E.C., Eickholt,B., Shaltiel,G., Agam,G., and Harwood,A.J. (2004c). A molecular cell biology of lithium. Biochem. Soc. Trans. *32*, 799-802.
- Winokur,G. and Reich,T. (1970). Two genetic factors in manic-depressive disease. Compr. Psychiatry *11*, 93-99.
- Wittekindt,O., Schwab,S.G., Burgert,E., Knapp,M., Albus,M., Lerer,B., Hallmayer,J., Rietschel,M., Segman,R., Borrmann,M., Lichtermann,D., Crocq,M.A., Maier,W., Morris-

Rosendahl,D.J., and Wildenauer,D.B. (1999). Association between hSKCa3 and schizophrenia not confirmed by transmission disequilibrium test in 193 offspring/parents trios. Mol. Psychiatry *4*, 267-270.

- Wittkopp,P.J., Haerum,B.K., and Clark,A.G. (2004). Evolutionary changes in cis and trans gene regulation. Nature *430*, 85-88.
- Wixon,J. (2000). Website review: UK CropNet. Yeast *17*, 244-254.
- Wong,A.H., Gottesman,I.I., and Petronis,A. (2005). Phenotypic differences in genetically identical organisms: the epigenetic perspective. Hum. Mol. Genet. *14 Spec No 1*, R11-R18.
- Woods,B.T. (1998). Is schizophrenia a progressive neurodevelopmental disorder? Toward a unitary pathogenetic mechanism. Am. J. Psychiatry *155*, 1661-1670.
- Wooster,R., Bignell,G., Lancaster,J., Swift,S., Seal,S., Mangion,J., Collins,N., Gregory,S., Gumbs,C., and Micklem,G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789-792.
- Workman,J.L. and Kingston,R.E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annu. Rev. Biochem. *67*, 545-579.
- Wray,G.A., Hahn,M.W., Abouheif,E., Balhoff,J.P., Pizer,M., Rockman,M.V., and Romano,L.A. (2003). The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. *20*, 1377-1419.
- Wu,B., Kitagawa,K., Zhang,N.Y., Liu,B., and Inagaki,C. (2004). Pathophysiological concentrations of amyloid beta proteins directly inhibit rat brain and recombinant human type II phosphatidylinositol 4-kinase activity. J. Neurochem. *91*, 1164-1170.
- Xiao,L. and Casero,R.A., Jr. (1996). Differential transcription of the human spermidine/spermine N1-acetyltransferase (SSAT) gene in human lung carcinoma cells. Biochem. J. *313 (Pt 2)*, 691-696.
- Xing,Q.H., Wu,S.N., Lin,Z.G., Li,H.F., Yang,J.D., Feng,G.Y., Wang,M.T., Yang,W.W., and He,L. (2003). Association analysis of polymorphisms in the upstream region of the human dopamine D4 receptor gene in schizophrenia. Schizophr. Res. *65*, 9-14.
- Xu,H., Sha,M.Y., Wong,E.Y., Uphoff,J., Xu,Y., Treadway,J.A., Truong,A., O'Brien,E., Asquith,S., Stubbins,M., Spurr,N.K., Lai,E.H., and Mahoney,W. (2003). Multiplexed SNP genotyping using the Qbead system: a quantum dot-encoded microsphere-based assay. Nucleic Acids Res. *31*, e43.
- Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B., and Kinzler,K.W. (2002). Allelic variation in human gene expression. Science *297*, 1143.
- Yang,J.Z., Si,T.M., Ruan,Y., Ling,Y.S., Han,Y.H., Wang,X.L., Zhou,M., Zhang,H.Y., Kong,Q.M., Liu,C., Zhang,D.R., Yu,Y.Q., Liu,S.Z., Ju,G.Z., Shu,L., Ma,D.L., and Zhang,D. (2003). Association study of neuregulin 1 gene with schizophrenia. Mol. Psychiatry *8*, 706-709.
- Yang,S. and Hecht,N.B. (2004). Translin associated protein X is essential for cellular proliferation. FEBS Lett. *576*, 221-225.
- Yang,X., Kuo,Y., Devay,P., Yu,C., and Role,L. (1998a). A cysteine-rich isoform of neuregulin controls the level of expression of neuronal nicotinic receptor channels during synaptogenesis. Neuron *20*, 255-270.
- Yang,X.L., Xiong,W.C., and Mei,L. (2004). Lipid rafts in neuregulin signaling at synapses. Life Sci. *75*, 2495-2504.
- Yang,Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555-556.
- Yang,Z., Boffelli,D., Boonmark,N., Schwartz,K., and Lawn,R. (1998b). Apolipoprotein(a) gene enhancer resides within a LINE element. J. Biol. Chem. *273*, 891-897.
- Ying,S.W., Futter,M., Rosenblum,K., Webber,M.J., Hunt,S.P., Bliss,T.V., and Bramham,C.R. (2002). Brain-derived neurotrophic factor induces long-term potentiation in intact adult hippocampus: requirement for ERK activation coupled to CREB and upregulation of Arc synthesis. J. Neurosci. *22*, 1532-1540.

- Yolken,R.H., Karlsson,H., Yee,F., Johnston-Wilson,N.L., and Torrey,E.F. (2000). Endogenous retroviruses and schizophrenia. Brain Res. Brain Res. Rev. *31*, 193-199.
- Yolken,R.H. and Torrey,E.F. (1995). Viruses, schizophrenia, and bipolar disorder. Clin. Microbiol. Rev. *8*, 131-145.
- Yoon,Y.R., Cha,I.J., Shon,J.H., Kim,K.A., Cha,Y.N., Jang,I.J., Park,C.W., Shin,S.G., Flockhart,D.A., and Shin,J.G. (2000). Relationship of paroxetine disposition to metoprolol metabolic ratio and CYP2D6*10 genotype of Korean subjects. Clin. Pharmacol. Ther. *67*, 567-576.
- Yoshida,K., Ito,K., Sato,K., Takahashi,H., Kamata,M., Higuchi,H., Shimizu,T., Itoh,K., Inoue,K., Tezuka,T., Suzuki,T., Ohkubo,T., Sugawara,K., and Otani,K. (2002). Influence of the serotonin transporter gene-linked polymorphic region on the antidepressant response to fluvoxamine in Japanese depressed patients. Prog. Neuropsychopharmacol. Biol. Psychiatry *26*, 383-386.
- Yoshimura,T., Kawano,Y., Arimura,N., Kawabata,S., Kikuchi,A., and Kaibuchi,K. (2005). GSK-3beta regulates phosphorylation of CRMP-2 and neuronal polarity. Cell *120*, 137-149.
- Young,S.T., Porrino,L.J., and Iadarola,M.J. (1991). Cocaine induces striatal c-fos-immunoreactive proteins via dopaminergic D1 receptors. Proc. Natl. Acad. Sci. U. S. A *88*, 1291-1295.
- Youngren,K.D., Inglis,F.M., Pivirotto,P.J., Jedema,H.P., Bradberry,C.W., Goldman-Rakic,P.S., Roth,R.H., and Moghaddam,B. (1999). Clozapine preferentially increases dopamine release in the rhesus monkey prefrontal cortex compared with the caudate nucleus. Neuropsychopharmacology *20*, 403-412.
- Yu,H., Peters,J.M., King,R.W., Page,A.M., Hieter,P., and Kirschner,M.W. (1998). Identification of a cullin homology region in a subunit of the anaphase-promoting complex. Science *279*, 1219-1222.
- Yuh,C.H., Bolouri,H., and Davidson,E.H. (1998). Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science *279*, 1896-1902.
- Yvert,G., Brem,R.B., Whittle,J., Akey,J.M., Foss,E., Smith,E.N., Mackelprang,R., and Kruglyak,L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. *35*, 57-64.
- Zanardi,R., Benedetti,F., Di Bella,D., Catalano,M., and Smeraldi,E. (2000). Efficacy of paroxetine in depression is influenced by a functional polymorphism within the promoter of the serotonin transporter gene. J. Clin. Psychopharmacol. *20*, 105-107.
- Zhang,X., Gainetdinov,R.R., Beaulieu,J.M., Sotnikova,T.D., Burch,L.H., Williams,R.B., Schwartz,D.A., Krishnan,K.R., and Caron,M.G. (2005). Loss-of-function mutation in tryptophan hydroxylase-2 identified in unipolar major depression. Neuron *45*, 11-16.
- Zhang,X.Y., Zhou,D.F., Zhang,P.Y., and Wei,J. (2000). The CCK-A receptor gene possibly associated with positive symptoms of schizophrenia. Mol. Psychiatry *5*, 239-240.
- Zhao,X., Shi,Y., Tang,J., Tang,R., Yu,L., Gu,N., Feng,G., Zhu,S., Liu,H., Xing,Y., Zhao,S., Sang,H., Guan,Y., St Clair,D., and He,L. (2004). A case control and family based association study of the neuregulin1 gene and schizophrenia. J. Med. Genet. *41*, 31-34.
- Zhao,X., Varnai,P., Tuymetova,G., Balla,A., Toth,Z.E., Oker-Blom,C., Roder,J., Jeromin,A., and Balla,T. (2001). Interaction of neuronal calcium sensor-1 (NCS-1) with phosphatidylinositol 4-kinase beta stimulates lipid kinase activity and affects membrane trafficking in COS-7 cells. J. Biol. Chem. *276*, 40183-40189.
- Zhou,Q.Y., Quaife,C.J., and Palmiter,R.D. (1995). Targeted disruption of the tyrosine hydroxylase gene reveals that catecholamines are required for mouse fetal development. Nature *374*, 640-643.
- Zhu,J., Liu,J.S., and Lawrence,C.E. (1998). Bayesian adaptive sequence alignment algorithms. Bioinformatics. *14*, 25-39.
- Zou,F., Li,C., Duan,S., Zheng,Y., Gu,N., Feng,G., Xing,Y., Shi,J., and He,L. (2005). A family-based study of the association between the G72/G30 genes and schizophrenia in the Chinese population. Schizophr. Res. *73*, 257-261.

APPENDIX

9 APPENDIX

9.1 Evaluation of scoring schemes for comparative genomics with vertebrate species

	HMR		HMRC			HMRZF				HMRCZF				
	mouse	rat	mouse	rat	chicken	mouse	rat	zebra	fugu	mouse	rat	chicken	zebra	Fugu
APOA2	0.6	0.6												
APOE	1.2	1.2												
ATP1A2	0.6	0.6												
COX6A2	0.9	1.2												
ENO3	0.6	0.5												
HP	1.1	1.1												
IL5	0.5	0.6												
OXT	1.4	1.2												
PGAM2	1.0	0.9												
TAGLN	0.5	0.5												
BIN1	0.5	0.5	0.5	0.5	2.3									
CHRND	0.7	0.8	0.7	0.8	3.4									
CHRNG	0.7	0.8	0.7	0.8	65.3									
CKMT2	0.5	0.7	0.5	0.7	1.6									
GNRH1	0.6	0.6	0.6	0.6	43.9									
LCP1	0.5	0.5	0.5	0.5	1.4									
MB	1.2	1.3	1.2	1.2	6.6									
MYL3	0.6	0.6	0.6	0.6	4.4									
NEFL	0.8	0.9	0.8	1.0	2.6									
PLAT	1.0	1.0	1.0	1.0	25.5									
SLC25A4	0.4	0.4	0.4	0.4	1.5									
ACTA1	1.2	1.4				1.2	1.3	2.8	8.6					
CHRNB1	0.6	0.7				0.6	0.7	52.1	57.9					
CHRNE	1.0	1.0				0.9	1.0	53.0	56.3					
CKM	0.8	0.9				0.8	0.9	2.7	2.7					
COL1A1	0.3	0.3				0.3	0.3	1.9	3.2					
DES	0.6	0.7				0.6	0.7	66.6	31.2					
EPO	0.7	0.8				0.9	0.9	1.1	1.7					
GP1BA	0.6	0.7				0.9	0.9	3.6	53.0					
CFTR	0.6	0.6	0.6	0.6	1.3	0.6	0.6	17.4	26.3	0.6	0.6	1.3	48.1	28.1
CRYAB	0.4	0.3	0.4	0.4	1.6	0.4	0.4	47.5	50.7	0.4	0.4	1.5	46.6	22.4
G6PC	0.7	0.8	0.7	0.7	57.8	0.7	0.7	2.6	4.4	0.7	0.7	56.1	2.4	5.3
HSPA5	0.4	0.4	0.4	0.4	1.3	0.4	0.4	12.8	6.9	0.4	0.4	1.3	13.0	7.7
IGF1	0.6	0.6	0.6	0.6	1.4	0.6	0.6	2.4	4.0	0.6	0.6	1.5	2.5	2.9
IGFALS	1.8	1.9	1.9	2.0	8.9	2.3	2.3	48.5	69.3	2.4	2.4	6.9	48.4	69.4
MYOG	0.5	1.0	0.6	0.8	16.2	0.5	1.0	30.8	70.7	0.6	0.8	14.7	17.0	54.1
PROC	1.1	1.4	1.1	1.4	18.5	1.2	1.6	20.6	4.5	1.2	1.6	18.0	29.1	4.6
RB1	0.4	0.4	0.4	0.4	1.6	0.5	0.6	50.6	20.5	0.5	0.5	1.6	37.5	20.8
STAR	0.9	0.7	0.9	0.7	3.5	0.9	0.7	2.9	2.5	0.9	0.7	3.3	3.3	2.6
TNNC1	0.9	0.8	0.9	0.8	4.7	0.9	0.8	3.0	3.8	0.9	0.8	4.7	3.0	3.7

Table S1: Rates of synonymous substitution. Estimates are given for the four sets of groups of orthologous genes: HMR: Human, mouse and rat; HMRC: Human, mouse, rat and chicken; HMRZF: Human, mouse, rat, zebrafish and fugu; HMRCZF: Human, mouse, rat, chicken, zebrafish and fugu. Each single column indicate the rate of synonymous substitution computed at four-fold positions.

			sensitivity				specificity				ppv			
Set	Score	Seq	av	sd	min	Max	av	sd	min	max	av	sd	min	max
HMR	bino	rr	30	17	1	69	96	3	89	100	26	16	3	65
HMRC	bino	rr	21	18	2	61	98	2	91	99	33	21	6	74
HMRZF	bino	rr	22	15	2	54	97	2	94	100	28	19	1	56
HMRCZF	bino	rr	26	19	7	59	98	1	96	100	40	22	8	77
HMR	diverg	rr	30	17	1	69	96	3	88	99	26	16	3	65
HMRC	diverg	rr	20	18	1	61	98	2	91	99	31	21	5	77
HMRZF	diverg	rr	22	14	2	50	97	2	93	100	29	16	5	56
HMRCZF	diverg	rr	22	12	7	45	98	2	94	99	35	19	9	70
HMR	disty	rr	30	17	1	72	96	3	88	99	26	16	3	65
HMRC	disty	rr	21	17	2	61	97	3	89	99	30	18	6	54
HMRZF	disty	rr	21	12	7	45	96	2	92	99	27	17	3	53
HMRCZF	disty	rr	19	14	7	45	97	2	92	99	29	18	4	52
HMR	pid	rr	30	17	1	72	96	3	88	99	26	16	3	65
HMRC	pid	rr	21	18	1	61	98	2	90	99	32	22	5	77
HMRZF	pid	rr	22	15	2	50	97	2	94	100	27	18	1	51
HMRCZF	pid	rr	25	14	7	45	98	2	95	99	37	19	13	70
HMR	bino	cds	70	17	3	93	96	3	89	100	66	18	2	95
HMRC	bino	cds	81	18	20	99	98	2	91	99	76	18	16	98
HMRZF	bino	cds	79	12	55	99	98	2	94	100	78	14	41	96
HMRCZF	bino	cds	81	13	55	93	98	1	96	100	78	16	44	97
HMR	diverg	cds	70	17	3	93	96	3	88	99	66	18	2	95
HMRC	diverg	cds	80	18	23	98	98	2	91	99	76	17	18	98
HMRZF	diverg	cds	78	14	50	99	97	2	93	100	75	16	42	96
HMRCZF	diverg	cds	79	15	54	93	98	2	94	99	75	17	41	96
HMR	disty	cds	71	17	8	94	96	3	88	99	66	18	4	95
HMRC	disty	cds	80	19	24	98	98	2	89	99	74	18	19	95
HMRZF	disty	cds	76	15	44	100	97	2	92	99	73	17	44	96
HMRCZF	disty	cds	77	15	53	92	98	2	92	99	73	18	40	91
HMR	pid	cds	71	17	8	94	96	3	88	99	66	18	4	95
HMRC	pid	cds	81	17	23	99	98	2	90	99	75	18	18	98
HMRZF	pid	cds	80	12	57	99	97	2	94	100	76	16	38	96
HMRCZF	pid	cds	80	13	54	94	98	1	95	99	76	17	41	96
HMR	bino	fct	57	18	7	83	96	3	89	100	69	17	6	95
HMRC	bino	fct	60	19	24	89	98	2	91	100	79	17	25	99
HMRZF	bino	fct	65	15	27	88	98	2	94	100	80	14	43	99
HMRCZF	bino	fct	66	19	26	88	98	1	96	100	81	15	49	99
HMR	diverg	fct	57	17	7	83	96	3	88	99	69	17	6	95
HMRC	diverg	fct	60	19	24	88	98	2	91	100	78	16	27	99
HMRZF	diverg	fct	63	16	27	88	97	2	93	100	77	16	42	96
HMRCZF	diverg	fct	63	18	26	84	98	2	94	100	78	17	44	99
HMR	disty	fct	58	17	11	83	96	3	88	99	69	17	6	95
HMRC	disty	fct	59	19	22	88	98	2	89	100	77	17	27	100
HMRZF	disty	fct	62	16	27	88	97	2	92	99	75	17	44	96
HMRCZF	disty	fct	61	19	26	83	98	2	92	100	76	19	42	99
HMR	pid	fct	58	17	11	83	96	3	88	99	69	17	6	95
HMRC	pid	fct	60	19	24	89	98	2	90	100	77	17	27	99
HMRZF	pid	fct	65	15	27	88	97	2	94	100	78	16	38	97
HMRCZF	pid	fct	65	18	26	85	98	1	95	100	79	17	46	99

Table S2 : Accuracy of detection of functional sequences. Data were obtained with a 20 bp window. 'HMR' : human, mouse and rat ; 'HMRC' : human, mouse, rat and chicken ; 'HMRZF' : human, mouse, rat, zebrafish and fugu ; 'HMRCZF' : human, mouse, rat, chicken, zebrafish and fugu ; 'bino': binomial score; 'diverg' score based on dS estimate, 'disty': score based on the distance in years between species; 'pid': score defined as the average percentage of identity; 'rr': regulatory region; 'cds': coding sequence; 'fct': functional sequence (i.e. 'rr' or 'cds'); ppv: positive predictive value ; 'av' : average ; 'sd' ; standard deviation ; 'min' : minimum ; 'max' maximum.

			sensitivity				specificity				ppv			
Set	Score	Seq	av	sd	min	max	av	sd	min	max	av	sd	min	max
HMR	bino	rr	31	18	10	69	97	3	89	100	37	22	5	82
HMRC	bino	rr	20	16	4	50	98	1	96	100	34	18	11	74
HMRZF	bino	rr	25	14	7	50	98	1	95	100	31	20	6	58
HMRCZF	bino	rr	24	18	1	50	98	2	96	99	27	17	4	51
HMR	diverg	rr	31	18	10	69	97	3	89	100	37	22	5	82
HMRC	diverg	rr	21	15	4	50	98	1	96	100	35	18	11	74
HMRZF	diverg	rr	26	15	8	50	97	2	94	99	36	17	5	54
HMRCZF	diverg	rr	22	16	1	44	97	2	94	99	23	13	4	41
HMR	disty	rr	31	18	10	75	97	3	87	100	37	22	5	82
HMRC	disty	rr	22	16	4	50	98	1	97	100	35	18	12	74
HMRZF	disty	rr	22	11	6	38	96	3	93	100	34	22	9	76
HMRCZF	disty	rr	18	15	1	44	97	2	93	99	20	11	4	39
HMR	pid	rr	31	18	10	75	97	3	87	100	37	22	5	82
HMRC	pid	rr	21	15	4	50	98	1	95	100	34	18	10	74
HMRZF	pid	rr	24	15	8	50	97	2	95	99	31	20	2	59
HMRCZF	pid	rr	26	21	1	63	98	2	95	99	26	15	4	44
HMR	bino	cds	76	17	6	98	97	2	89	100	74	18	4	99
HMRC	bino	cds	82	19	24	98	98	3	85	100	76	18	19	96
HMRZF	bino	cds	82	9	65	97	98	1	95	100	78	12	49	95
HMRCZF	bino	cds	80	13	56	92	98	2	95	99	74	15	45	91
HMR	diverg	cds	76	17	6	98	97	2	89	100	74	18	5	99
HMRC	diverg	cds	82	18	30	98	98	3	85	100	76	17	23	96
HMRZF	diverg	cds	80	11	57	97	97	2	94	99	76	14	46	95
HMRCZF	diverg	cds	78	15	52	92	97	2	94	99	71	17	45	90
HMR	disty	cds	77	17	6	98	97	3	87	100	74	18	4	99
HMRC	disty	cds	81	18	30	98	98	3	85	100	76	17	24	96
HMRZF	disty	cds	79	12	55	97	97	2	93	100	74	15	46	95
HMRCZF	disty	cds	76	16	46	92	97	2	93	99	69	19	33	90
HMR	pid	cds	77	17	6	98	97	3	87	100	74	18	4	99
HMRC	pid	cds	82	18	30	98	98	3	85	100	76	17	24	96
HMRZF	pid	cds	81	9	65	97	97	2	95	99	77	14	42	95
HMRCZF	pid	cds	78	14	52	92	97	2	94	99	72	16	45	90
HMR	bino	fct	61	18	10	87	97	2	89	100	76	18	9	101
HMRC	bino	fct	60	20	24	88	98	3	85	100	79	18	27	103
HMRZF	bino	fct	66	15	26	89	98	1	95	100	80	12	52	98
HMRCZF	bino	fct	63	19	24	85	98	2	96	99	76	14	50	92
HMR	diverg	fct	61	18	10	88	97	2	89	100	76	18	9	100
HMRC	diverg	fct	60	19	24	88	98	3	85	100	78	17	30	103
HMRZF	diverg	fct	64	16	25	89	97	2	94	100	78	14	46	98
HMRCZF	diverg	fct	61	19	24	83	97	2	94	99	74	15	50	90
HMR	disty	fct	61	19	10	88	97	3	87	100	76	18	9	101
HMRC	disty	fct	60	20	23	87	98	3	85	100	78	17	31	103
HMRZF	disty	fct	63	16	24	86	97	2	93	100	76	15	46	103
HMRCZF	disty	fct	59	19	24	82	97	2	93	99	72	18	39	90
HMR	pid	fct	61	19	10	88	97	3	87	100	76	18	9	101
HMRC	pid	fct	60	19	24	88	98	3	85	100	78	17	31	103
HMRZF	pid	fct	66	15	25	89	97	2	95	100	78	14	44	98
HMRCZF	pid	fct	62	19	24	83	97	2	95	99	75	14	52	90

Table S3 : Accuracy of detection of functional sequences. Data were obtained with a 40 bp window. 'HMR' : human, mouse and rat ; 'HMRC' : human, mouse, rat and chicken ; 'HMRZF' : human, mouse, rat, zebrafish and fugu ; 'HMRCZF' : human, mouse, rat, chicken, zebrafish and fugu ; 'bino': binomial score; 'diverg' score based on dS estimate, 'disty': score based on the distance in years between species; 'pid': score defined as the average percentage of identity; 'rr': regulatory region; 'cds': coding sequence; 'fct': functional sequence (i.e. 'rr' or 'cds'); ppv: positive predictive value ; 'av' : average ; 'sd' ; standard deviation ; 'min' : minimum ; 'max' maximum.

			Sensitivity				specificity				ppv			
Set	Score	Seq	av	sd	min	max	av	sd	min	max	av	sd	min	max
HMR	bino	rr	35	22	4	75	97	3	90	99	38	26	7	81
HMRC	bino	rr	31	21	3	59	98	1	96	99	36	19	3	59
HMRZF	bino	rr	30	15	6	50	97	2	94	99	31	20	12	73
HMRCZF	bino	rr	30	24	4	59	97	2	94	99	26	17	12	53
HMR	diverg	rr	36	23	4	79	97	3	89	99	38	26	7	81
HMRC	diverg	rr	29	20	3	59	98	1	95	99	33	16	3	50
HMRZF	diverg	rr	28	13	6	50	97	2	94	99	30	19	11	73
HMRCZF	diverg	rr	38	32	18	74	98	2	96	99	33	12	20	44
HMR	disty	rr	36	23	4	79	97	3	89	99	38	26	7	81
HMRC	disty	rr	27	18	3	50	98	1	95	99	31	16	3	46
HMRZF	disty	rr	23	11	6	37	96	2	94	99	33	20	13	64
HMRCZF	disty	rr	38	32	18	74	98	2	96	99	32	13	19	44
HMR	pid	rr	36	23	4	79	97	3	89	99	38	26	7	81
HMRC	pid	rr	31	21	6	59	98	1	96	99	35	17	5	55
HMRZF	pid	rr	28	13	6	50	97	2	94	99	32	21	9	73
HMRCZF	pid	rr	33	37	6	74	98	1	97	99	28	22	12	53
HMR	bino	cds	80	16	15	99	97	2	90	99	75	16	12	98
HMRC	bino	cds	76	20	11	98	97	5	78	99	69	19	10	95
HMRZF	bino	cds	77	15	37	97	97	3	89	99	73	16	30	94
HMRCZF	bino	cds	72	18	28	92	97	2	94	99	66	20	22	88
HMR	diverg	cds	80	16	15	99	97	2	89	99	75	16	12	98
HMRC	diverg	cds	75	20	11	98	97	5	78	99	68	19	10	95
HMRZF	diverg	cds	78	15	37	97	97	2	93	99	73	16	27	94
HMRCZF	diverg	cds	71	17	28	92	96	2	93	99	64	18	24	87
HMR	disty	cds	80	16	15	99	97	2	89	99	75	16	12	98
HMRC	disty	cds	75	20	11	98	96	5	78	99	67	19	10	90
HMRZF	disty	cds	76	15	37	97	97	2	92	99	70	17	26	94
HMRCZF	disty	cds	71	17	28	92	96	2	93	99	63	19	22	86
HMR	pid	cds	80	16	15	99	97	2	89	99	75	16	12	98
HMRC	pid	cds	76	20	11	98	97	5	78	99	69	19	10	95
HMRZF	pid	cds	78	15	37	97	97	2	93	99	74	17	24	94
HMRCZF	pid	cds	72	18	28	92	96	2	93	99	64	19	24	88
HMR	bino	fct	61	18	21	88	97	2	90	99	76	15	20	98
HMRC	bino	fct	56	20	6	84	97	5	78	100	71	20	10	95
HMRZF	bino	fct	62	16	22	86	97	3	89	100	74	16	36	101
HMRCZF	bino	fct	56	20	21	85	97	2	94	99	68	19	29	90
HMR	diverg	fct	61	17	21	88	97	2	89	99	76	15	20	98
HMRC	diverg	fct	55	20	6	84	97	5	78	100	70	19	10	95
HMRZF	diverg	fct	63	16	22	87	97	2	93	100	75	16	33	101
HMRCZF	diverg	fct	55	18	18	78	96	2	93	99	66	17	36	87
HMR	disty	fct	62	17	21	88	97	2	89	99	76	16	20	98
HMRC	disty	fct	55	20	6	84	97	5	78	100	69	19	10	95
HMRZF	disty	fct	60	16	22	86	97	2	92	100	72	17	26	99
HMRCZF	disty	fct	54	18	18	78	96	2	93	99	66	18	34	86
HMR	pid	fct	62	17	21	88	97	2	89	99	76	16	20	98
HMRC	pid	fct	56	20	6	85	97	5	78	99	71	19	10	95
HMRZF	pid	fct	63	16	22	86	97	2	93	100	75	16	29	101
HMRCZF	pid	fct	55	18	18	78	96	2	93	99	66	17	36	90

Table S4 : Accuracy of detection of functional sequences. Data were obtained with a 100 bp window. 'HMR' : human, mouse and rat ; 'HMRC' : human, mouse, rat and chicken ; 'HMRZF' : human, mouse, rat, zebrafish and fugu ; 'HMRCZF' : human, mouse, rat, chicken, zebrafish and fugu ; 'bino': binomial score; 'diverg' score based on dS estimate, 'disty': score based on the distance in years between species; 'pid': score defined as the average percentage of identity; 'rr': regulatory region; 'cds': coding sequence; 'fct': functional sequence (i.e. 'rr' or 'cds'); ppv: positive predictive value ; 'av' : average ; 'sd' ; standard deviation ; 'min' : minimum ; 'max' maximum.

9.2 Published papers

9.2.1 The meso-genomic era

(Semple et al., 2001)

9.2.2 SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis

(Le Hellard et al., 2002)

http://genomebiology.com/2001/2/7/reports/4015.1

Meeting report

The meso-genomic era

Colin AM Semple, Martin S Taylor and Stephane Ballereau

Address: Department of Medical Sciences, Molecular Medicine Centre, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.

Correspondence: Colin AM Semple. E-mail: Colin.Semple@ed.ac.uk

Published: 28 June 2001

Genome ***Biology*** 2001, **2(7)**:reports4015.1–4015.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/7/reports/4015

A report from HGM2001, the sixth annual International Human Genome Meeting organized by The Human Genome Organisation (HUGO), Edinburgh, UK, 19-22 April 2001.

In his opening remarks, the president of HUGO, Lap Chee Tsui, included a plea to avoid the term 'post-genomic era'. One does not have to look hard to see why. According to the Genome Monitoring Table [http://www.ebi.ac.uk/genomes/mot/] hosted by the European Bioinformatics Institute (EBI) about 45% of the human genome is available in finished form, and the rest is available as fragmented 'draft' sequence. Also, accurate and comprehensive descriptions of the functions and often even the structures of the 30,000-40,000 predicted genes are, at the moment, rare. Added to this, the catalog of variations seen in the human genome is growing but far from complete. At the same time, the Entrez Genome pages [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome] at the National Center for Biotechnology Information (NCBI) say that there are currently sequences for 9 finished or draft eukaryotic, 40 bacterial and 10 archaean genomes, but a veritable avalanche of new genomes is already on its way. This new wave of data will extend genomics to new protozoan, fungal, plant and animal taxa, including eagerly awaited vertebrates such as zebrafish, pufferfish, the chimp and other apes. From several perspectives we are decidedly in the 'mid-genomic' era.

At this conference a good sampling of activities in the present 'meso-genomic' era was presented and several themes recurred. We carried out a superficial survey of the abstracts available at the HGM2001 website [http://hgm2001.hgu.mrc.ac.uk/] and found certain terms to be better represented than others (Figure 1). Many abstracts (134 out of the 543 online) dealt with studies of human variation and particularly of single-nucleotide polymorphisms (SNPs). Far fewer abstracts contained references to human diseases or disorders, reflecting the fact that much work on variation was concerned simply with the detection of SNPs and issues around data analysis. In such 'data-rich times' computational analysis is very much in evidence. Much of what we know about the function of human genes is inferred computationally and so, to rectify this, studies are underway to generate functional data using model organisms. Comparative genomics is rapidly coming of age and many studies, particularly those using the draft mouse genome, demonstrated its utility in sequence annotation. Fewer studies at this meeting concentrated on what precisely is transcribed and eventually translated from the genome, although important initiatives were described. It is also becoming evident that epigenetic mechanisms can be important in the journey from genotype to phenotype.

The finer details: genomic variation and complex traits

It is often quoted that all humans are around 99.9% genetically identical, but that still leaves millions of sites in the genome where we differ. The largest public repository for SNPs, dbSNP [http://www.ncbi.nlm.nih.gov/SNP/] at the NCBI, now contains more than 1.68 million of these variations. At the HUGO conference it became clear that the search is on for the variations that are important in identifying predisposition to disease and tracing our evolutionary history. A key question, which received a lot of attention, is the extent of linkage disequilibrium (LD) between SNPs. LD is a measure of the association between alleles; for example if no recombination has occurred between alleles they are in complete LD. The greater the distance such associations span, the more chance we have of finding SNPs that indicate predisposition to disease. A great deal depends on the ancient history of human populations. Some previous theoretical work, assuming constant population expansion, has suggested that LD should extend over only a few kilobases. It now appears, however, that the history of Northern European populations involved bottleneck events since migration

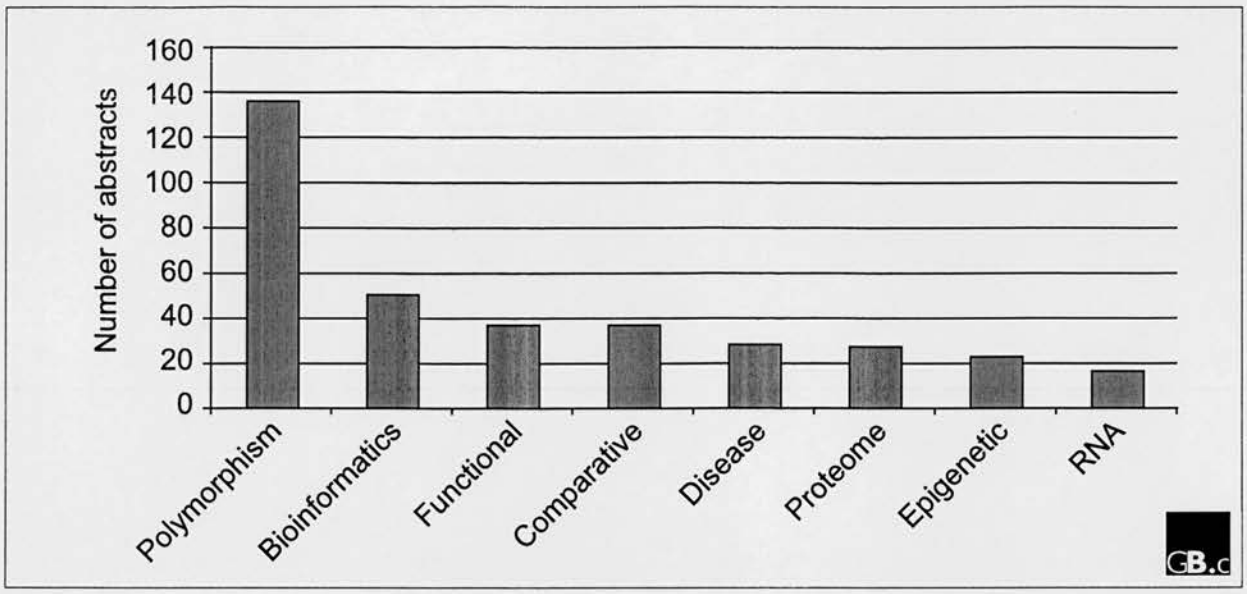


Figure 1
A survey of common terms in abstracts submitted to HGM2001. The categories on the horizontal axis were derived from searches for the presence of the following terms: polymorphism* or SNP*; bioinformatic* or comput*; function* and genom*; compar* and genom*; disease* or disorder*; proteom* or protein* and structur*; epigenet* or methylat* or heterochrom*; RNA and splic* or process*. Individual abstracts can appear in more than one category, so little can be concluded from the absolute heights of columns.

from Africa. Eric Lander (Whitehead Institute, Cambridge, USA) revealed that this has resulted in LD extending over some regions of around 120 kb (although great variation in LD was seen across the genome) and that a map defining perhaps 30,000 ancestral regions of LD can be constructed for such populations. In a Nigerian population, by contrast, LD extended over less than 10 kb and is more consistent with a simple model of population expansion. The results suggest that a representative set of only 100,000 SNPs could be effective for whole-genome association studies of disease in Northern European populations. Michael Olivier (Stanford Human Genome Center, USA) found LD extending over regions of chromosome 21 that were between 2 and 85 kb, but these regions were disrupted by segments of similar or larger size that showed no significant LD between SNPs. In addition, each segment was characterized by a restricted number of observed haplotypes, with the commonest haplotype found in over 50% of all chromosomes. These results illustrate the need to characterize the pattern of LD in a region of interest before association studies are undertaken. They also imply that initial localization of disease loci may be easier in populations with LD that extends over more than 100 kb (such as Northern European populations), but that finer mapping of such loci, to specific areas of genes, may be easier in populations with smaller blocks of LD.

Joseph Terwilliger (Columbia University, New York, USA) discussed the feasibility of identifying common susceptibility alleles that exert small effects. Much depends upon the relationships between three factors: the marker genotype, the susceptibility allele and the phenotype of interest. In association studies, we measure the correlation between the marker genotype and the phenotype. The success of this enterprise depends on two relationships: between the marker genotype and the susceptibility allele (measured by linkage and/or LD), and between the susceptibility allele and the phenotype. The latter relationship depends in turn on the capacity to infer the genetic factor from the phenotype (detectance). With complex traits such as predisposition to common diseases, however, there may be many different causal genotypes, making detectance low. In such cases, linkage and LD mapping may not give significant results. All this still leaves aside the question of gene-environment interactions, which can further complicate the observed patterns of disease.

A model association study was presented by Xavier Estevill (Duran i Rynals Hospital, Barcelona, Spain), who investigated susceptibility to panic and phobic disorders. His group identified an interstitial duplication on chromosome 15q24-26 that is significantly associated with panic disorder within families from a small Catalonian village. The duplication is estimated to be present in around 7% of the general population. The association was then replicated in a more diverse population of patients from elsewhere in Spain and followed up with studies of the expression and function of

http://genomebiology.com/2001/2/7/reports/4015.3

genes within the duplication. One of these genes was overexpressed in transgenic mice and was found to cause an enhanced panic reaction in behavioral tests. In addition, the duplication was found in different forms within families, showed an absence of linkage with other 15q markers and exhibited mosaicism in patients' cells. It therefore appears to represent a novel, non-Mendelian disease mechanism. Allen Roses (GlaxoSmithKline, Triangle Park, USA) discussed successes in association studies based on SNPs. High-density disequilibrium mapping of SNPs has identified new susceptibility genes for psoriasis, migraine, diabetes mellitus (type 2) and Parkinson's disease within previously defined large linkage regions. He predicted that within the next few years clinical tests for drug efficacy or hypersensitivity will emerge using collections of SNPs that are associated with a given phenotype. The successes in identifying associations and the emerging LD patterns in the human genome have renewed optimism among those working on common diseases that show complex inheritance patterns.

The conference also included talks covering complex traits in species other than human. Complex genetic traits have been studied and manipulated in agriculture for thousands of years. The conventional breeding schemes of 'breeding the best to the best' have been incredibly successful, to the extent that most people would not think of eating - or indeed even recognizing - wild tomatoes. Dani Zamir (Hebrew University of Jerusalem, Israel) presented a model system for investigating quantitative trait loci (QTL) in tomato. Zamir and co-workers produced nearly isogenic lines (NILs) of tomato but included exotic chromosome segments from wild breeds of tomato, essentially using a back-cross strategy. Armed with NILs covering the entire genome, they found and mapped 23 QTLs for the brix (sugar) content of the fruit. Extension of the original strategy focused on one of the strongest QTLs, narrowing it to 484 bp within a gene for an apoplastic invertase, a key enzyme in sugar metabolism. Further work demonstrated that lower levels of mRNA from this gene are found in commercial cultivars than in the wild tomato. Not only is the tomato a good model for studying QTLs, it is also of substantial commercial significance: higher brix levels mean better ketchup.

Michel Goeorges (University of Liege, Belgium) described the fascinating inheritance pattern of "beautifully proportioned buttocks" (callipyge), in sheep. A ram with the desired trait was initially identified; through breeding of this ram and by using a least exclusion breakpoint mapping approach (that is, using recombination breakpoints to identify a common haplotype between individuals), a single, apparently autosomal dominant locus was pinpointed on chromosome 18 with a logarithm of odds (LOD) of 55. Classical genetic crosses were carried out and these revealed imprinting at the locus, with only those offspring that inherited a paternal callipyge allele expressing the phenotype. Back-crossing to produce sheep homozygous at the callipyge locus revealed that only heterozygous animals with a paternally derived allele expressed the phenotype, a phenomenon described as polar overdominance. Within the callipyge-determining locus four genes have been found, all of which are imprinted. It will be interesting to see the molecular explanation for this bizarre pattern of inheritance.

The same but different: comparative genomics

The human genome is unfinished and we still lack another complete vertebrate genome for comparison, but comparative genomics is already becoming popular. Jean Weissenbach (National Centre for Scientific Research (CNRS), Paris, France) described successes in comparisons between the human and more compact vertebrate genomes using the draft sequences from the two pufferfish species *Fugu rubripes* and *Tetraodon nigroviridis*. His group found such comparisons to be a valuable addition to the output of *ab initio* gene-prediction programs such as GenScan, for detecting novel exons and conserved, non-coding regions assumed to be important in regulation. By comparison, they had found human versus mouse comparisons to be more sensitive (detecting a larger number of exons) but also 'noisier' (generating many uninformative matches). Several studies were presented that exploit the draft mouse sequence to aid the annotation of the human genome. Lisa Stubbs (Lawrence Livermore National Laboratory, Livermore, USA) described a large-scale comparison between human chromosome 19 and 15 syntenic regions of the mouse genome adding up to 46 Mb of mouse sequence within 35 bacterial artificial chromosome (BAC) contigs. Combining these comparisons with other annotation methods (*ab initio* predictions and identifying matches to expressed sequences) gave a total of around 1,200 genes for human chromosome 19, with perhaps 50 genes identified by comparative genomics that had been missed by other methods. The mouse comparisons also identified many new candidate exons, 5′ untranslated regions (UTRs) and more than 4,000 non-coding conserved regions. Around 30% of the genes on human chromosome 19 are members of tandemly duplicated clusters, including vomeronasal receptors, olfactory receptors and zinc-finger genes. The regions containing clusters were found to have diverged considerably between human and mouse, varying in gene content, number and organization, in contrast to other regions.

Peter Holland (University of Reading, UK) has investigated the evolution of vertebrate genomes by studying the numbers of homeotic-gene clusters in vertebrates and amphioxus (the living invertebrate that is closest to the vertebrates). He postulates two episodes of tetraploidy, followed by selective gene loss in early chordate evolution near the time point at which the vertebrates emerged. Gazing further back in evolutionary time, Eugene Koonin (NCBI, Bethesda, USA) described using the combination of analyses of protein structure and comparative genomics to discover ancient protein domains originating in the last universal

common ancestor (LUCA) of all extant life forms. The main biological functions of the LUCA appear to have revolved around RNA metabolism and translation, whereas the DNA processing machinery is a later innovation. Koonin also offered insights into the way new domains evolved, by duplication and accretion, during eukaryotic evolution. Many eukaryote-specific domains are the result of explosive bursts of innovation; for example, exploiting the basic α helix in a variety of novel configurations such as the coiled coil.

It is clear that comparative genomics has thrown up a variety of new challenges for bioinformatics. Comparisons between multiple, long sequences that may only share small regions of limited homology is, in itself, a non-trivial problem. Burkhard Morgenstern (Munich Information Center for Protein Sequences (MIPS), Germany) presented a new version of the DIALIGN program that performs well in the identification of small regulatory regions in large non-coding DNA sequences. But although many conference participants emphasized the success of combinations of methods in gene prediction, we still lack software that combines intrinsic sequence properties (as detected by *ab initio* methods) with genomic-sequence comparisons and analyses of similarity to proteins or expressed sequence tags (ESTs). Tim Hubbard (Sanger Centre, Hinxton, UK) described how the Ensembl [http://www.ensembl.org] database, which provides annotation of the draft human genome, has incorporated comparisons to the draft mouse sequence. Hubbard and colleagues are now grappling with the prospect of incorporating comparisons between human and several other vertebrate genomes.

In spite of substantial conservation between species, there are clearly very significant differences in the genes and/or their regulation between species. One such difference between species is being investigated by Svante Pääbo (Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany) using comparative genomics. His group focuses on the differences between humans and other, closely related species. Previously, the only well-characterized biochemical difference between humans and great apes was the level of hydroxylation of sialic acids (major components of cell surfaces in animals), which is of major importance for the success of xenotransplantation. Comparing the relative expression levels of 20,000 genes in blood, liver and brain between humans, chimpanzees and rhesus macaques has revealed that the rate of 'transcriptome' change has been similar between chimps and humans (using macaques as a standard) in blood and liver. In stark contrast, the rate of change of expression patterns in the brain is accelerated approximately three-fold in humans. A number of the most strikingly different expression patterns are now being followed up.

The wild frontier: from genotype to phenotype

With an established foothold in genomics and large-scale, public collections of microarray data on gene expression just around the corner, people have begun to peer over the horizon at the proteome. Most human genes have not been characterized functionally, so what we do know about them mostly comes from computational detection of homology. As Janet Thornton (University College London, UK) made clear, protein family members commonly exhibit some similarity in function, but there is remarkable functional variation within many families. By examining enzymes within 167 structural superfamilies from the CATH database [http://www.biochem.ucl.ac.uk/bsm/cath/CATH.html], which classifies groups of proteins by class, architecture, topology or homology, she estimated that only around 90 superfamilies contained members with conserved functions. In variable superfamilies, members often differed in substrate specificity. Thus, although structural homology can extend computational predictions beyond homology at the sequence level, there will always be many proteins for which no reliable predictions can be made. Thornton also discussed an analysis of proteins associated with disease garnered from the Online Mendelian Inheritance in Man (OMIM) [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM] database. By superimposing disease-associated mutations onto these protein structures she concluded that there was no general pattern in the positions of disease-causing mutations. Some occurred within regions, such as active sites, that would be predicted to alter function, but many others appear to relate to interactions with other proteins. It is widely accepted that a wider sampling of the protein-structural universe is needed to enhance our understanding of protein function and dysfunction. Tim Harris (Structural Genomix, San Diego, USA) discussed large-scale approaches to X-ray crystallography which might transform information obtained from the human sequence into novel three-dimensional protein structures. Initially, the focus is on structural families with pharmaceutical significance.

Traditionally, the development of new drugs involves testing novel compounds in disease models with little knowledge of the biochemical pathways or physiological systems underlying the disease. Because disruptions in the abundances and activities of proteins are the molecular basis of many diseases, proteomics promises to revolutionize drug discovery by allowing rational identification of drug targets. But the set of proteins present within a cell at any one time is the net outcome of many complex processes, including synthesis, degradation, modification and processing. Each protein can then participate in myriad interactions with other molecules in different places within the cell. In addition, protein abundances, cellular locations, activities and interactions may differ across tissues and between normal and disease states. A number of approaches to investigate this awe-inspiring complexity were presented. Ian Tomlinson (MRC Laboratory for Molecular Biology, Cambridge, UK) described recombinant-antibody technologies for use in large-scale studies of protein localization, quantification and interactions. The aim is to construct high-density antibody arrays

http://genomebiology.com/2001/2/7/reports/4015.5

that are able to profile the expression of most human proteins in a given tissue within the next 3-10 years. Mark Vidal (Dana-Faber Cancer Institute, Boston, USA) outlined his approach to building a protein-interaction map for the *Caenorhabditis elegans* proteome. The idea is to amplify and clone every open reading frame in the genome for use in extensive yeast two-hybrid system assays. The resulting interaction map is to form part of an envisaged functional atlas for *C. elegans*. As Patricia Kuwabara (Sanger Centre, Hinxton, UK) pointed out, because around three quarters of human disease genes have *C. elegans* homologs, the worm could be a valuable source of functional annotation for the human genome. She outlined efforts to carry out comprehensive DNA microarray expression profiling as well as RNA-mediated interference (RNAi) the latter of which provides a rapid, sequence-specific method to abolish gene activity.

The mouse continues to be a favorite model organism for investigating gene function, and in this field investigations are also increasing in the form of screens and multiple studies carried out in parallel. Keats Nelms (John Curtin School of Medical Research, Canberra, Australia) and Sally Cross (MRC Human Genetics Unit, Edinburgh, UK) described the use of the mutagen N-ethyl-N-nitrosourea (ENU) for the generation of thousands of mice with random point mutations throughout their genome. All of the mutagenesis screens are picking up dominant, highly penetrant morphological and behavioral phenotypes. Each project is also following up specific groups of phenotypes: mutants affecting the immune system in the case of Keats Nelms and mutants affecting eye development and eye disease with Sally Cross. Whereas most ENU screens of mice are designed to identify only dominant mutations, the strategy employed by Nelms and co-workers generates pedigrees to uncover recessive mutations and intrinsically provides the basic resources to map the locus of the mutation by back-crossing. Not surprisingly, 90% of the mutations that have been found are recessive. As an alternative strategy for uncovering recessive ENU-induced mutations, Allan Bradley (Sanger Centre, Hinxton, UK) has been coordinating the development of mouse lines with *Cre-lox*-induced megabase deletions, which, when crossed with ENU-mutagenized mice would uncover recessive mutations in a locus-targeted manner. Other innovations from the Bradley lab include coat-color chromosome tags and megabase inversions, both of which have the potential to speed up the characterization of mutant mice.

The armored, three-spined stickleback (*Gasterosteus aculeatus*) is no stranger to the laboratory, having been well characterized in terms of morphology, color and behavior. Now David Kingsley (Stanford University, USA) and colleagues are using these fish to characterize the molecular genetic changes that are responsible for the morphological evolution of vertebrates. Geographically isolated populations of stickleback have undergone substantial adaptive radiation in lakes left in the wake of retreating glaciers around 10,000 years ago. The different species vary in size, shape, behavior, color and patterns of defensive armor. As most of these species have undergone parapatric divergence (as a result of population subdivision and reproductive isolation), isolating mechanisms are largely behavioral and can be overcome in the laboratory to produce fully viable F1 and F2 generations. With cDNA-library resources and a genome-wide linkage map in place, Kingsley is crossing species pairs and watching the divergent traits segregate in the offspring. It is immediately apparent that many of the segregating morphologies are caused by genes that have a major effect, rather than by many genes that have a relatively minor effect. These loci are being mapped and eventually the molecular basis of the trait will be uncovered. For one of the armor-patterning traits, it has already been found that one locus has independently changed twice to produce the same morphological adaptation.

Wendy Bickmore (MRC Human Genetics Unit, Edinburgh, UK) opened the workshop on chromosome structure and epigenetic mechanisms with the observation that it is often assumed that by combining the genome sequence with information about gene expression and the proteome we will understand inheritance. This is wrong, she said, because epigenetic processes such as histone modification and DNA methylation as well as chromatin-remodeling systems can all be important. A striking example was given by Mark Bailey (University of Glasgow, UK) who described his studies of the methyl CpG-binding protein 2 gene (MECP2). Mutations in this gene cause Rett syndrome, a severe X-linked, neurodevelopmental disorder that is the commonest cause of severe cognitive incapacity in the female population. MECP2 is believed to bind methylated CpG islands and mediate transcriptional repression. Bailey's work has identified genomic fragments to which MECP2 binds strongly. Intriguingly, many of these fragments contain Alu interspersed repeat elements, suggesting the possibility that Alu overexpression could be a cause of Rett syndrome. Not only the pioneering work going on in epigenetic mechanisms but also many other studies presented at the meeting showed that even when we finally have a functionally annotated human genome there will still be much to learn.

Nucleic Acids Research, 2002, Vol. 30 No. 15 e74

SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis

Stéphanie Le Hellard[1,*], Stéphane J. Ballereau[1], Peter M. Visscher[2], Helen S. Torrance[1], Jeni Pinson[1], Stewart W. Morris[1], Marian L. Thomson[1], Colin A. M. Semple[1,3], Walter J. Muir[4], Douglas H. R. Blackwood[4], David J. Porteous[1] and Kathryn L. Evans[1]

[1]Medical Genetics Section, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK, [2]Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK, [3]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK and [4]Department of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK

Received March 14, 2002; Revised and Accepted May 27, 2002

ABSTRACT

We have compared the accuracy, efficiency and robustness of three methods of genotyping single nucleotide polymorphisms on pooled DNAs. We conclude that (i) the frequencies of the two alleles in pools should be corrected with a factor for unequal allelic amplification, which should be estimated from the mean ratio of a set of heterozygotes (k); (ii) the repeatability of an assay is more important than pinpoint accuracy when estimating allele frequencies, and assays should therefore be optimised to increase the repeatability; and (iii) the size of a pool has a relatively small effect on the accuracy of allele frequency estimation. We therefore recommend that large pools are genotyped and replicated a minimum of four times. In addition, we describe statistical approaches to allow rigorous comparison of DNA pool results. Finally, we describe an extension to our ACeDB database that facilitates management and analysis of the data generated by association studies.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common type of polymorphism in the human genome, with an approximate frequency of one every kilobase (1). These biallelic variants are relatively easy to genotype compared with VNTRs and microsatellites. For these reasons SNPs are thought to have a promising future in a wide range of human genetics applications including pharmacogenomics, the study of population evolution, analysis of forensic samples and the identification of susceptibility genes involved in complex diseases. Hence, a large proportion of the effort of genome centres is now focused on the identification and the mapping of a large collection of SNPs: to date about 1 260 000 have been mapped onto the human draft sequence (http://snp.cshl.org/).

The study of complex common diseases and quantitative traits is confounded by the effects of disease heterogeneity, gene–gene and gene–environment interactions. This means that large numbers of SNPs must be surveyed in large numbers of individuals in order to detect single gene variants with a small to moderate effect size (2,3). The use of pooled samples, comprised of equal amounts of genomic DNA from up to 1000 individuals, has been proposed as a means of reducing the number of genotyping reactions required. The method used to genotype SNPs in pooled DNAs must provide accurate estimates of allele frequencies, and must be time and cost effective. The spectra of methods currently available for genotyping SNPs in individual samples [for an extensive review of SNP genotyping methods see Syvanen (4)] can be divided into three classes. First, methods such as SSCP or dHPLC that are based on the physical–chemical properties of the alleles. Secondly, methods such as TAQMAN™ (Applied Biosystems); oligo-ligation assay; Invader assay™ (Third Wave Technologies Inc.); and allele-specific amplification and padlock probes that are based on hybridisation, amplification or ligation of an allele-specific probe. Thirdly, methods based on allele-specific extension or minisequencing from a primer adjacent to the site of the SNP such as SNaPshot™ (Applied Biosystems); primer extension read by dHPLC or by mass spectrometry; primer extension performed on microarrays; fluorescence polarisation; bioluminometric assay coupled with modified primer extension reactions (BAMPER) and Pyrosequencing™ (Pyrosequencing).

Previous studies have shown that allelic frequencies can be accurately estimated from pools using primer extension followed by dHPLC (5); TAQMAN™ and RFLP analysis (6); allele-specific amplification with real-time PCR (7); SSCP (8); BAMPER (9) and MassARRAY™ (10). In common with many other groups, we wish to screen a large candidate region

*To whom correspondence should be addressed. Tel: +44 131 651 1061; Fax: +44 131 651 1059; Email: s.lehellard@ed.ac.uk

e74 Nucleic Acids Research, 2002, Vol. 30 No. 15

for evidence of genetic association. The preferred strategy is to assay small numbers of pooled DNA samples with large numbers of SNPs. Consequently, methods such as Pyrosequencing™, TAQMAN™ or BAMPER that use modified primers are too expensive. Methods based on hybridisation or on physical–chemical properties are ruled out as each assay must be optimised. We therefore chose to compare the robustness, accuracy and cost of three methods based on minisequencing: SNaPshot™ (Applied Biosystems) and primer extension followed either by dHPLC, or mass spectrometry (MassARRAY™ system by Sequenom).

We have also addressed the important issues of how many DNAs can be pooled, and how many times pool genotypes should be replicated to optimise the accuracy of allele frequency estimation.

In addition, we suggest the use of a modified statistical method that allows rigorous analysis of allele frequencies estimated from pools. Classical association studies on individual DNA samples use the χ^2 test to compare the frequencies of alleles in case and control populations. However, when pooled DNAs are used, allelic frequencies are estimated rather than directly counted from individual genotypes, which introduces extra sources of error. We have therefore modified the χ^2 test to take these sources of error into account, diminishing the risks of type I error.

Finally, genotyping large numbers of SNPs on pools or on individual samples generates a large data set. We have set up an extension of our ACeDB database (11) to store and manage information on the pools, individuals and markers and to record and analyse genotyping results. Furthermore, we have created in ACeDB a model ('Pop_pool_meta') that allows the data of several pools or populations of individual samples to be merged and analysed as a single set. This option allows the pools or populations to be stratified on the basis of phenotypic traits, and then analysed independently or together. We have also developed a 'user friendly' web interface for submission of new data, which is fed automatically into an analysis pipeline, before being recorded in the database.

MATERIALS AND METHODS

DNA pool set up

All subjects gave written ethical consent to take part in these studies.

The concentration of the DNAs used to construct pools was measured using the PicoGreen dsDNA Quantitation Reagent (Molecular Probes) in a CytoFluor fluorimeter (Applied Biosystems). The DNAs were diluted to a final concentration of 8 ng/µl and equal amounts of DNAs were mixed to form the pools.

Range pools were constructed by mixing appropriate volumes of homozygote DNA. Five homozygote DNAs for each genotype were used for pooling. The concentrations ranged from 50–50% to 85–15%, with 5% increments.

Markers and PCR

SNPs RS643304, RS1402045, RS15020285, RS489009 and RS508509 were retrieved from the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP). Primers were designed using Primer3 programme (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi), and synthesised by Genosys Biotechnologies (Europe) Ltd. Sequences of PCR and genotyping primers are available upon request.

PCRs were carried out on a PTC225 (MJ Research) using 40 ng total DNA, 10 pmol of each primer, 80 µM dNTPs (Sigma), 1.5 mM $MgCl_2$ and 1 U *Taq* (Sigma) in 1× PCR buffer [67 mM Tris–HCl, 16 mM enzyme grade $(NH_4)_2SO_4$, 6.7 mM $MgCl_2$ pH 8.8]. The programme used was an initial denaturation of 94°C for 3 min, followed by 10 cycles of 94°C for 15 s, touch down annealing from 65 to 55°C for 30 s over 10 cycles (–1°C/cycle) and 72°C for 45 s, followed by 30 cycles of 94°C for 15 s, 55°C for 30 s and 72°C for 45 s.

PCR clean up

After PCR, the products were checked on a 2% agarose gel. PCR primers and dNTPs were removed before genotyping: 5 µl of PCR product was incubated with 1 µl of ExoSapIT (Amersham Pharmacia) for 45 min at 37°C, followed by 20 min at 80°C for enzyme inactivation. For multiplexing of PCRs, 1 µl of each PCR product was pooled and treated with 1 µl of ExoSapIT.

Primer extension followed by dHPLC

Reactions were carried out as described in Hoogendoorn *et al.* (12).

SNaPshot™ reactions

The primers used for the extension reactions were designed according to the manufacturer's recommendations. Additionally, we used the mfold server (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi) to assess the secondary structure of the PCR product and the accessibility of the SNP, in order to decide whether to use the forward or the reverse primer.

Reactions were carried out in a final volume of 10 µl, containing 2 µl of cleaned up PCR product, 1 µl of SNaPshot™ multiplex mix (Applied Biosystems), 2 µl of half term buffer (200 mM Tris–HCl, 5 mM $MgCl_2$ pH 9), 2 pmol of genotyping primer. In multiplex reactions 2 µl of the cleaned PCR multiplex was used. The cycling programme was 25 cycles of 94°C for 10 s, 50°C for 5 s, 60°C for 30 s. After cycling, the unincorporated fluorescent ddNTPs were removed by adding 1 U of shrimp alkaline phosphatase (Amersham Pharmacia) and incubating for 45 min at 37°C, followed by 20 min at 80°C for enzyme inactivation. An aliquot of 9 µl formamide was added to 1 µl of SNaPshot™ reactions and loaded on ABI3700 sequencer (Applied Biosystems). Samples were run using the POP6 Polymer, with dye set E and analysed using the Genescan v3.5.2 program. The relative proportion of each allele was measured by the height of the corresponding peaks.

Primer extension on MassARRAY™

Reactions were performed at Sequenom GmbH (Hamburg, Germany; http://www.sequenom.com). Assays were designed using Sequenom's SpectroDESIGNER™ software (version 1.3.4). Genotypes were performed using MassARRAY™ system and SpectroTYPER™ software.

Statistical analysis

Correction for unequal allelic amplification and estimation of frequencies. Let k be the ratio of the two allele peak heights in heterozygotes. Following Hoogendoorn *et al.* (5), this factor is estimated from a number of independent heterozygotes, and we assume that the estimator $\hat{k}$ is unbiased with a variance of σ_k^2, i.e. $\hat{k} \sim (k, \sigma_k^2)$ with $\sigma_k = SE(\hat{k})$. The error in estimating k arises from variation in the quality of the DNA from each heterozygote, and from a pure experimental error attached to each individual analysis. The estimate of the allele frequency in the pool is estimated as

$$\tilde{p} = A/(A + \hat{k}B),$$

with $\tilde{p}$ the estimated frequency in pools, and A and B the observed peak heights corresponding to the two alleles. The variance of the estimated allele frequency as a function of the variance of $\hat{k}$ is, approximately, following a Taylor series expansion,

$$\text{var}(\tilde{p} \text{ due to } \hat{k}) \approx p^2 (1 - p)^2 CV^2(\hat{k})$$

with p the true frequency in the population, and CV the coefficient of variation. Furthermore we observed a pool specific error (e) which contributes to a difference between the allele frequency estimated from the pools and the estimate of the allele frequency from a direct count of alleles on individual genotypes ($\hat{p}$),

$$\tilde{p} = \hat{p} + e$$

We assumed that these errors are normally distributed. This assumption was confirmed for the distribution of the frequencies for 10 replicates of the five markers we have tested in this study. Following these assumptions, the variance of the estimated allele frequency from a pool of N individuals is

$$\begin{aligned}\text{var}(\tilde{p}) &\approx \text{var}(\hat{p}) + \text{var}(e) + \text{var}(\tilde{p} \text{ due to } \hat{k}) \\ &= p(1 - p)/(2N) + \text{var}(e) + p^2 (1 - p)^2 CV^2(\hat{k})\end{aligned}$$

Comparing freqencies between pools. The standard procedure to test whether the allele frequencies in two pools are significantly different from each other is to summarise the observed counts in a 2×2 table and to perform a χ^2 test (13). For a case-control study we use the following notation,

	cases	controls	
Allele 1	a	b	$a + b = N_{all1}$
Allele 2	c	d	$c + d = N_{all2}$
	$a + c = N_{case}$	$b + d = N_{control}$	N

In this notation N_{case} is twice the number of case individuals and, for an equal number of cases (n) and controls (n), $N = 2n + 2n = 4n$. The standard test statistic of independence can be written as

$$T_1 = (ad - bc)^2 N/(N_{case} \times N_{control} \times N_{all1} \times N_{all2})$$

Under the null hypothesis of the same population allele frequencies in cases and controls, for large N and not too extreme population frequencies, this test is distributed as a χ^2 with one degree of freedom. If estimated counts are substituted for the observed ones, this test is then T_{est}

$$T_{est} = (a_e d_e - b_e c_e)^2 N/(N_{case} \times N_{control} \times N_{all1} \times N_{all2})$$

with a_e, b_e, c_e and d_e the estimates of a, b, c and d, respectively. The expected value of T_{est} is, approximately,

$$E(T_{est}) \sim 1 + \text{var}(e)/[2\text{var}(\hat{p}_0)],$$

with $\hat{p}_0$ the estimate of the allele frequency across the two pools under the null hypothesis, i.e. $\hat{p}_0 = (a + b)/N$, and its variance is obtained from the binomial distribution [$\text{var}(\hat{p}_0) = \hat{p}_0(1 - \hat{p}_0)/N$]. Under the null hypothesis of equal allele frequencies, the expected value of the test statistic based upon observed counts is $E(T_1) = 1$. Hence, the test statistic is inflated by the extra source of errors in estimating the allele frequencies and its use would lead to an inflated type I error rate. We suggest a simple adjusted test,

$$T_{adj} = T_{est} \times [2\text{var}(\tilde{p}_0)]/[2\text{var}(\tilde{p}_0) + \text{var}(e)],$$

i.e. a shrunk version of the standard test statistic, with the estimate of the sampling variance of the allele frequency under the null hypothesis obtained from the estimated counts (i.e. $\tilde{p}_0$ replacing $\hat{p}_0$).

A more detailed protocol is available online at our website (http://www.genetics.med.ed.ac.uk/protocols/).

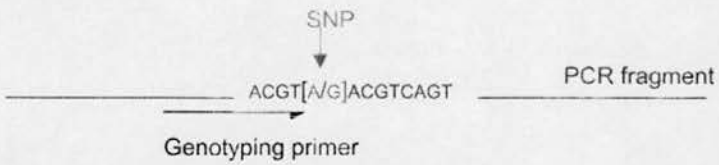
Data management in ACeDB

Modified models. The 'STS' model was modified to cross-reference to the allele model for the markers (SNPs or microsatellites) that are present in the STS. The 'allele' model was modified to store information on the method used to genotype the marker (e.g. details of the extension primer and genotyping method used, and the experimental conditions); the pools, populations, metapopulations and metapools that have been typed with the marker and the statistical analysis of the results obtained.

New models. The 'Individual' model was created to store individual genotypes and indicate which pool or population, metapopulation and metapool an individual DNA sample belongs to.

The new 'Pop_pool' model stores the identities of the DNA samples that constitute the pool or the population, and provides links to the markers that have been genotyped on the pool/population and the genotyping results. It also stores the results of comparison of allele frequencies with those of other pools/populations, or metapools/metapopulations, and information regarding inclusion of the pool/population in a metapool/metapopulation.

The 'Pop_pool_metapool' model stores data on the set of pools/populations combined and provides links to the markers typed on the pools/populations, the statistical description of the data set obtained, and information on the comparison of allele frequencies with those of other samples (pools, populations, metapopulations and metapools).



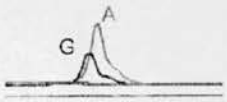
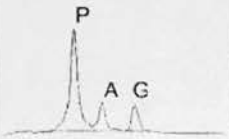
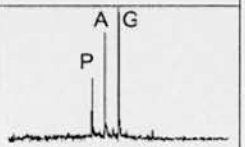
	SNaPshot™	primer extension-dHPLC	MassARRAY™
Reaction mix (enzyme, buffer)	Labelled ddNTPs: ddATP ddCTP ddGTP ddTTP	only one ddNTP e.g. ddTTP, dCTP	ddTTP and dCTP
Extension product for allele A	→ ddT	→ ddT	→ ddT
Extension product for allele G	→ ddC	→ CddT	→ CddT
Output for an heterozygote P: unextended primer A: peak for allele A G: peak for allele G			

Figure 1. Genotyping a SNP with SNaPshot™, primer extension followed by dHPLC or mass spectrometry (MassArray™) analysis. The three methods are based on the allele specific extension of a genotyping primer adjacent to the SNP site (see below). The region containing the primer is first amplified by PCR.

The statistical description model 'Stat_des' stores the results of the statistical analysis of data obtained by genotyping specified markers on a given sample (mean frequencies of the two alleles, standard deviation, standard error).

Finally, the statistical comparison model 'Stat_comp' stores the results of association studies carried out by comparing allele frequencies in the samples or direct allele counts for a given marker, using the appropriate statistical test.

Web interface for submission of data and automation of statistical analysis

CGI/Perl scripts were produced to facilitate the submission of new data and to perform statistical description and χ^2 tests. When entering new data or updating a given object, the tace program (Morris, J. 1994; http://www.acedb.org/Cornell/tace.html) is used by the script to retrieve existing information, e.g. names of the pool and the marker, and information on the genotyping experimental conditions. The user then selects parameters from pull-down lists that are either defined in the script or retrieved from the database (e.g. names of the pool and the marker), which ensures accurate data entry. Pool genotyping data are then automatically analysed using a script that computes descriptive statistics and runs the Shapiro–Wilk test for goodness-of-fit to normality.

Tests of association based on pools and metapools are automatically performed using a script that runs the modified χ^2 test. The interface also allows classical association studies to be carried out based on genotypes of individuals. A '.ace' file storing any new data submitted is generated and read into the database. All the models and scripts described here are available at http://www.genetics.med.ed.ac.uk.

RESULTS

Figure 1 gives a brief description of the methods compared in this study.

Estimating allele frequency in pools: correction for unequal allele amplification

By definition, heterozygote individuals have an equal number of copies of the two alleles at any given locus. If genotyping was equally efficient for the two alleles, then the two amplified peaks would be the same height. However, in practice unequal peak heights is the norm. We genotyped individual heterozygotes 6–10 times and recorded the variation in peak height

Table 1. Unequal allelic amplification

Marker	SNaPshot™	dHPLC	MassARRAY™
RS15020285	0.951 ± 0.272 (0.086)	0.416 ± 0.272 (0.087)	0.566 (nd)
RS508509	0.268 ± 0.085 (0.027)	0.389 ± 0.016 (0.005)	0.593 (nd)
RS1402045	0.447 ± 0.047 (0.015)	0.911 ± 0.465 (0.147)	0.754 (nd)
RS643304	0.648 ± 0.009 (0.003)	0.748 ± 0.063 (0.020)	0.564 (nd)
RS489009	0.813 ± 0.275 (0.087)	0.569 ± 0.161 (0.051)	0.515 (nd)

Ten heterozygote individuals were genotyped for each of the five markers. The mean ratio of the two allele peaks across 10 different heterozygotes individuals ± standard deviation between the ratios of the heterozygotes (and standard error of the mean) is shown. nd, not determined.

ratios between replicates. This indicated that the standard error of the mean (SEM) ratios were less than one-twentieth of the mean value (data not shown).

We observed greater variation in peak height ratios between different heterozygotes (SEM ranging from 0.003 to 0.147; Table 1). Hoogendoorn *et al.* (5) reported a SEM of 0.005–0.06 when comparing the peak height ratios of different heterozygotes for nine markers genotyped by primer extension followed by dHPLC. It is possible that the variation observed between heterozygotes could be due to variable DNA quality. However, all of the DNAs used for this study were collected, extracted and stored under the same conditions. The observed variation is therefore more likely to be caused by factors specific to the experimental procedures. The maximum variation (SEM) between the ratios of different heterozygote individuals was within one-seventh of the mean ratio (Table 1). A mean ratio of unequal amplification (k) can therefore be accurately calculated for any given marker genotyped by a given method.

When allele frequencies are estimated by genotyping a pooled sample the resultant peak heights must be corrected for unequal amplification by the factor k. If k is the mean ratio of the allele A and B peak heights (H_A and H_B, respectively) in heterozygote individuals, i.e. $k = H_A/H_B$, then the frequency of the allele A in the pools would be $p_{Apools} = H_{Apools}/(H_{Apools} + kH_{Bpools})$. If the unequal amplification of alleles is ignored and the ratios of peak heights are used in a χ^2 statistic this will result in a biased test procedure, as the test statistic is not distributed as a χ^2 under the null hypothesis of equal allele frequencies in the pools. To a first order approximation, the expected value of the test statistic based upon the unadjusted ratio of peak heights is

E(test statistic) = k/[1 + (k – 1)p]**2,

where p is the population frequency. This result was validated by computer simulation. Depending on the true value of k and the frequency (p), this test is either expected to be smaller or larger than 1.0 (which is the expected value from a proper χ^2 test). For example, for p = 0.25 and k = 0.5, the expected value of the test statistic is ~0.65, which will result in a test that is too conservative.

Hoogendoorn *et al.* (5) used the mean ratio from eight heterozygote individuals to determine k. We currently use a panel of 16 control individuals, which are genotyped for each marker. We then calculate a mean ratio from the heterozygote individuals of this panel—we always divide the height of the smaller allele peak by the height of the bigger allele peak to keep data homogenous. As long as the SEM of the ratio between the heterozygotes is less than one-tenth of the mean value, we use this mean value as k. If the SEM is greater than one-tenth of k, then either the assay must be optimised or the number of heterozygote genotypes must be increased appropriately until this criterion is met.

To avoid introducing any extra sources of variation, we perform all stages of the genotyping procedure simultaneously on all samples.

Effect of allele frequencies on unequal amplification in pools

The allele frequencies in pools of cases and controls are both corrected with the same k factor. This approach is valid only if unequal amplification is linearly correlated with allele frequencies. To test this, we constructed two sets of artificial pools with a range of allele frequencies by mixing appropriate volumes of two homozygote DNAs for the markers RS1402045 and RS643304. The ratio of concentrations of alleles in both sets of pools ranged from 50–50% to 85–15%, at 5% increments. Each pool was genotyped 5–10 times. Results (Fig. 2) show that there is a linear correlation between the allele frequencies and the ratios for all three methods tested. These data indicate that variation in allele frequency does not affect the extent of unequal allele amplification, and that pools with different allele frequencies can be corrected with the same k factor.

Comparison of accuracy and repeatability of SNaPshot™, primer extension followed by dHPLC, or by mass spectrometry (MassARRAY™) methods

Five markers were genotyped on a set of 96 individual DNAs to obtain the sample allelic frequencies. The 96 DNAs were then pooled (Pool96), and the pool was genotyped 10 times with each of the five markers by the three methods. The k factors were obtained for the three methods and used to correct the estimate of allele frequencies within the pools. The estimated allele frequencies were in good agreement with the results of individual genotyping. This was true for all methods and markers tested (Table 2). Several parameters are important in comparing the efficiency of the different methods. First, the estimation of frequencies has to be close to the sample frequencies, as a large discrepancy would introduce a risk of type I and type II errors. However, as we demonstrate below, good repeatability (i.e. a smaller SEM) is more important than pinpoint accuracy. Poor repeatability necessitates a larger

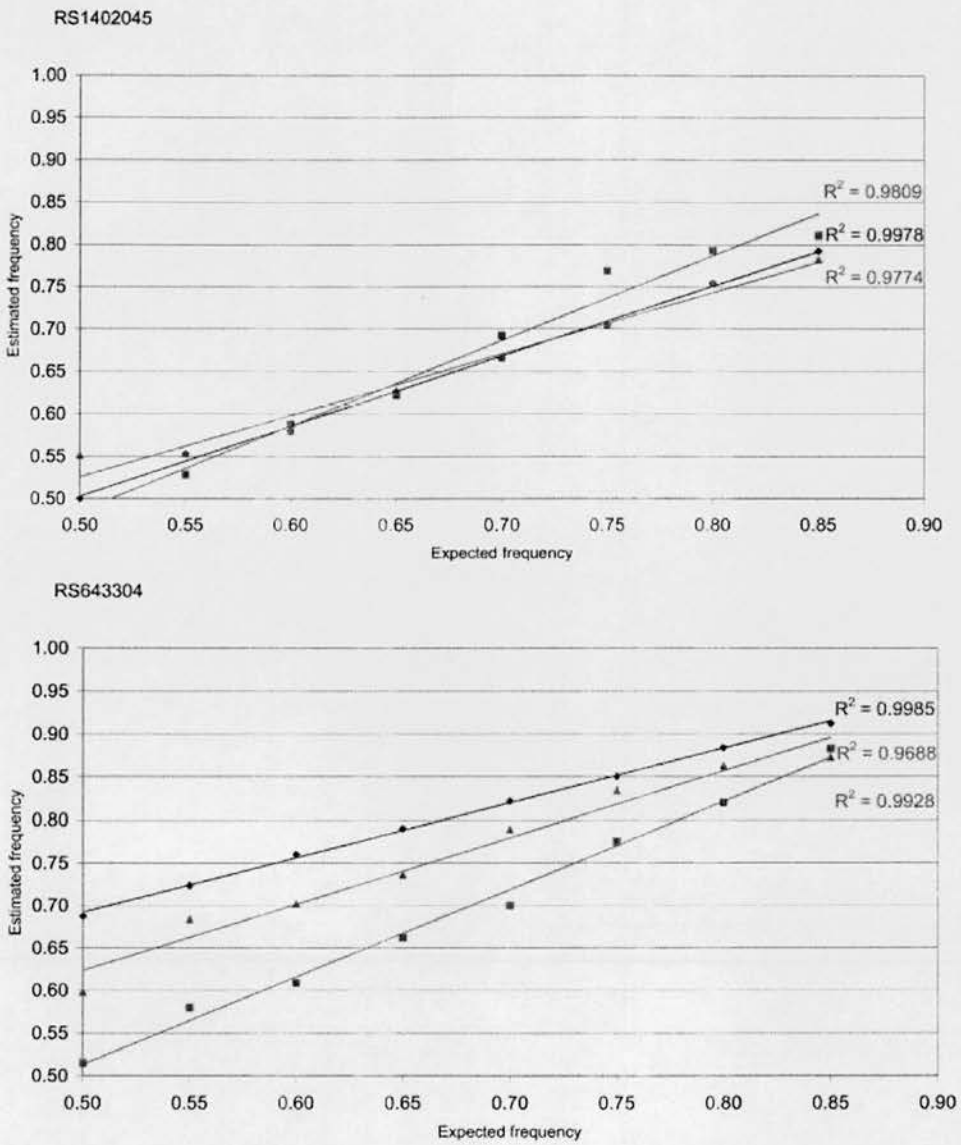


Figure 2. Test for linearity for the markers RS1402045 and RS643304 across artificial pools with allele frequencies ranging from 50–50% to 85–15% with 5% increments were constructed. Comparison of the three methods: SNaPshot™ (diamonds), dHPLC (triangles) and MassARRAY™ (squares).

Table 2. Comparison of the accuracy of the three methods in estimating allele frequencies in a pool of 96 DNAs (Pool96)

Marker	Sample frequency	Estimated frequency		
		SnaPshot™	dHPLC	MassARRAY™
RS15020285	0.658	0.666 ± 0.022 (0.007)	0.633 ± 0.013 (0.004)	0.727 ± 0.013 (0.004)
RS508509	0.714	0.702 ± 0.135 (0.043)	0.711 ± 0.013 (0.004)	0.761 ± 0.009 (0.003)
RS1402045	0.713	0.699 ± 0.047 (0.015)	0.683 ± 0.047 (0.015)	0.724 ± 0.009 (0.003)
RS643304	0.657	0.648 ± 0.032 (0.010)	0.656 ± 0.022 (0.007)	0.645 ± 0.013 (0.004)
RS489009	0.528	0.561 ± 0.063 (0.02)	0.501 ± 0.054 (0.017)	0.503 ± 0.009 (0.003)

The sample allelic frequencies were obtained from genotyping the 96 individuals (using the SNaPshot™ method). A k correction factor for unequal amplification was obtained for each of the three methods and used to estimate the frequencies in the Pool96. Ten replicates of Pool96 were genotyped to test the repeatability of the method, which is expressed here as ± standard deviation (and SEM).

Table 3. Comparison of cost and throughput of the methods

	Cost per sample	Throughput/day/machine
dHPLC	157 cents	192
MassARRAY™	100 cents	40 000
SNaPshot™	87 cents	7000[a]
		1500[b]

The calculations include genotyping reagents and primers (on the basis of 100 reactions per primer), plastic consumables and the cost of running the assay on the detection platform. Salary costs are not included. PCR costs are not included, as they are the same for the three methods.
[a]With a 96 capillary system (ABI37000).
[b]With a 16 capillary system (ABI3100).

number of replications in order to lower the SEM of the estimated allele frequencies. The SEMs observed varied from 0.003 to 0.066 for SNaPshot™, 0.003 to 0.017 for primer extension followed by dHPLC and 0.003 to 0.004 for MassARRAY™. Thus, from a quantitative point of view, the three methods tested are all suitable for genotyping pools, with the MassARRAY™ method performing substantially better than the other two.

Ease of use and cost considerations

For all of the markers tested, the SNaPshot™ method was found to be robust and required little optimisation. However, multiplexing SNP assays was less straightforward, as the signal strength varied between assays. We circumvented this problem by multiplexing assays on the basis of signal strength, or by increasing the amount of genotyping primer for the weaker assays (14). We currently find that multiplexing four markers is relatively straightforward, although according to the manufacturer 10 SNPs can be successfully multiplexed.

In our hands, SNP genotyping using primer extension followed by dHPLC required extensive optimisation of the primer extension reaction. Optimisation of the gradient that is best suited to the elution of each product was also required and, furthermore, attempts to multiplex reactions were unsuccessful.

For the MassARRAY™ analysis, sequence files with information on the marker and localisation of the SNPs to be detected were provided to Sequenom who designed the assays using their 'in house' software. Coded samples and pools were provided, and highly satisfactory results returned promptly for each SNP assay.

The cost of genotyping pools (Table 3) is highly dependent on the ability to multiplex reactions and minimise reaction volume. For these reasons, primer extension followed by dHPLC or MassARRAY™ is not as cost effective as the SNaPshot™ method. However, as we demonstrate below, the MassARRAY™ requires less replicates per pool than the SNaPshot™, which makes MassARRAY™ as cost effective as SNaPshot™. Table 3 also provides a comparison of the throughputs of the different platforms.

Effect of pool size

We wanted to determine whether the number of the samples in the pool would affect the accuracy of allele frequency estimation in pooled DNAs. We genotyped 384 individual DNAs to calculate the sample frequencies for the markers RS1402045, RS15020285 and RS643304. The same 384 individuals were then included in four pools of 96, two pools of 192, one pool of 288 and one pool of 384 individuals, and genotyped using the SNaPshot™ method (six replicates per pool). The frequencies estimated from the pools were compared with the sample frequencies (Fig. 3). We found that for the range tested, pool size had no significant effect on the accuracy of frequency estimations or on repeatability. This indicates that pooling larger numbers of samples does not result in a loss of power. We therefore recommend that larger pools are typed, minimising the number of genotyping reactions required.

Statistical comparison of two or more pools genotyped with a given marker

Estimating the allelic frequencies in pools. As we have demonstrated above, estimating allele frequencies in pooled DNAs introduces three potential sources of error: (i) error caused by sampling a finite number of individuals from a population (the standard sampling error); (ii) error in estimating the adjustment factor k; and (iii) a pool-specific measurement error. The first source of error is reduced by increasing the sample size; the second source of error is reduced by using the appropriate number of heterozygotes to estimate k (see above); and the third source of error is reduced by genotyping replicate samples of the pools. When allele frequencies from two (or more) pools are compared, a minor error in the estimation of k will induce a covariance between the estimates from the two pools, because the error in estimating k is the same for both pools. However, as the same error in estimating k is made for both pools, and as k is independent of the allelic frequency, the difference between the estimates of the frequency in both pools is not affected to a first order approximation. This was also observed by Hoogendoorn *et al.* (5).

Comparing frequencies between pools. We have modified the standard χ^2 test, which is used in classical case-control association studies, to take into account the sources of error discussed above. The effect of a larger variance in allele frequencies from pools due to the use of estimated rather than observed counts was investigated using models that simulate observed and estimated counts. The results are shown in Tables 4 and 5. Generally, unless the sources of errors are large, the inflation in the type I error is small. However, if the pool-specific error is large [e.g. experimental error (σ_e) > 0.025], then the type I error can be substantially inflated. For example, for $\sigma_e = 0.025$ the type I error is at least doubled relative to the type I error rate on the observed counts.

Regarding the type II error, power is reduced when using the adjusted statistical test relative to the power based upon observed counts (Table 5). For $\sigma_e > 0.025$, the reduction in power can be substantial. To achieve the same power for pooling and direct genotyping, the pool sample size must be increased by a factor of $1/[1 - 2\text{var}(e)/\text{var}(\Delta)]$, with var(Δ) the variance of the difference in allele frequencies in the two groups obtained from observed counts, and var(e) the experimental pool-specific error. For example, for $\sigma_e = 0.01$ (which corresponds to a standard error of 0.01, as is typically seen in SNaPshot™ experiments) and $\sigma_\Delta = 0.03$ (which

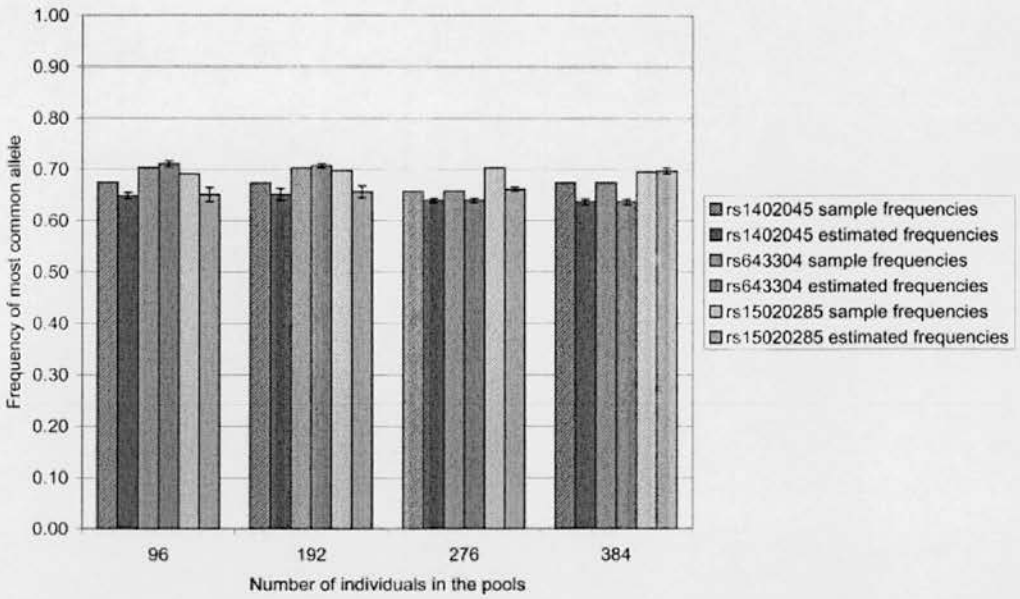


Figure 3. Estimation of allele frequency in different sized pools. Marker RS1402045, RS15020285 and RS643304 were typed on 384 individuals (using the SNaPshot™ method) to obtain sample allelic frequencies. 384 DNAs were combined in four pools of 96, two pools of 192, one pool of 276 and one pool of 384 individuals. Each pool was genotyped six times (using the SNaPshot™ method) and the frequencies were estimated from the mean frequency corrected for unequal amplification. The repeatability is expressed as the SEM estimated frequency.

Table 4. Empirical type I errors from 10 000 simulations, for 100 cases and 100 controls, and p = 0.5

		Using observed counts	Using estimated counts	
α^a	$\sigma_e{}^b$	T_1	T_{est}	T_{adj}
0.10	0.01	0.099	0.113	0.093
	0.025	0.098	0.180	0.094
	0.05	0.100	0.345	0.102
0.05	0.01	0.051	0.060	0.048
	0.025	0.051	0.112	0.053
	0.05	0.051	0.264	0.054
0.01	0.01	0.011	0.015	0.011
	0.025	0.011	0.039	0.011
	0.05	0.011	0.144	0.011

[a]Nominal type I error.
[b]SEM of estimated allele frequency.
T_1 χ^2 test on observed counts.
T_{est} unadjusted test on estimated counts.
T_{adj} adjusted test on estimated counts.

Table 5. Power for a significance level of 0.05 and 100 cases and 100 controls, from 10 000 simulations

$\sigma_e{}^a$	p(cases)	p(controls)	$T_1{}^b$	$T_{adj}{}^c$
0.01	0.50	0.45	0.17	0.16
		0.40	0.52	0.48
		0.35	0.86	0.83
0.025	0.50	0.45	0.17	0.13
		0.40	0.52	0.38
		0.35	0.86	0.71
0.05	0.50	0.45	0.17	0.09
		0.40	0.52	0.22
		0.35	0.86	0.42

[a]SEM of estimated allele frequency.
[b]T_1 χ^2 test on observed counts.
[c]T_{adj} adjusted test on estimated counts.

corresponds to, for example, 200 case and 200 control individual populations with frequencies of 0.3 and 0.2, respectively), the sample size of the pool would have to be increased by a factor of $1/(1 - 0.0002/0.0009) = 1.3$.

To achieve an experimental error of $\sigma_e = 0.01$ or less, replicate pools must be used. If the estimate of the between-replicate variation in the estimate of the allele frequency is in the range of 0.02–0.04 (standard deviation), then to achieve a SEM of <0.01, approximately 4–16 replicate pools would give the same power as tests based upon observations, assuming that there are no errors in determining individual genotypes. From the results in Table 2 we can conclude that most standard deviations are in this range, so that a minimum of four replicates appears to be appropriate.

Database development

We have previously used an ACeDB database (15) to manage the construction of a physical map of chromosome 4p16.1–15.3 (16). Although this database necessitates expert bio-informatics support, it possesses the flexible architecture required to adapt it to our current purpose. We were able to modify existing models and create new ones to allow storage of all information relevant to pool construction, populations and genotyping results. The new models facilitate storage of statistical analysis and of association data based on both pools

and populations of individual samples. We have also created a model ('Pop_pool_meta') that allows the genotyping data obtained for a set of pools or populations to be merged and analysed as a single data set. The data from different pools can then be merged and the results analysed as a single group.

Submission and analysis of new genotyping data

Data submission via a web interface. CGI/Perl scripts were produced to facilitate the submission of new data and to perform direct statistical analysis. A web page with a graphical representation of a 96-well plate allows the submission of individual genotypes for use in classical association studies. This interface can also be used to enter peak heights obtained from heterozygote samples, which are used to calculate the correction factor k. Another form is used to submit the allele peak heights for the different pool replicates.

Statistical analysis and description of pool genotyping data. Once entered, peak height ratios are subjected to statistical analyses (the mean frequencies of the two alleles, standard deviation and SEM are calculated). The interface also allows the analysis of a data set produced by combining two or more pools. A table of results is displayed and a file containing the new data and their statistical description is automatically created and read into the database.

Association studies based on pools. The modified χ^2 test is used to detect differences between the allele frequencies of selected pools. The results are displayed and stored in a new file that is then read into the database.

Association studies based on populations. For a specified marker, allele numbers in each selected population, or group of populations, are calculated from the genotypes of all individuals. A classical χ^2 test is performed on these data to identify differences between populations. Results are displayed and automatically read into the database.

DISCUSSION

Pool allele frequencies can be estimated with a high degree of accuracy using SNaPshot™ and primer extension followed by either dHPLC or MassARRAY™. However, accurate estimation of allele frequencies requires calculation of a correction factor for unequal allelic amplification from the peak height ratios of a small set of heterozygotes. Of the three methods tested, the MassARRAY™ method gives the best repeatability, while primer extension followed by dHPLC requires more optimisation than the other methods and does not allow easy multiplexing. The number of samples in a pool has a negligible effect on the accuracy of frequency estimations. We therefore recommend the use of larger pools (we use pools of 384 individuals) and multiple replicates rather than smaller pools with fewer replicates. The ideal number of replicates required is dependent on the reliability of the marker, and the repeatability of the method. For example, for the majority of markers four replicates appeared to be sufficient when the MassARRAY™ method was used.

Choosing a method for genotyping, particularly if this implies the purchase of expensive equipment, is difficult and no golden rule can be applied. The deciding factors include the number of genes/SNPs to be typed, the need for single genotyping versus pool genotyping, the level of throughput required and whether there is a need for SNP detection as well as genotyping. For example, MassARRAY™ may be the best choice for a core facility that provides a very high throughput SNP genotyping service on pools and/or individuals, but a capillary electrophoresis instrument would provide more flexibility for a project that requires SNP detection and medium genotyping throughput.

We have expanded the ACeDB architecture to allow the storage, management and analysis of genotyping data and related information. The ACeDB database provides a graphical, multi-window interface and allows the user to navigate easily between objects. With the new models one can now navigate from a marker to the pools tested with the marker, to the data obtained, and to the results of the comparison of allele frequencies in that pool with those of other pools. Additionally, we have developed a web interface that allows easy and accurate submission of new data and their automatic examination, via descriptive statistical analysis and association studies. A useful future development of the analysis pipeline would be a graphical display of the association data along the DNA sequence of a genomic region. This would allow researchers to visualise the strength of the association in the context of other sequence annotation.

We have modified an existing statistical test to correct for extra sources of error introduced by the pooling methodology. Our test controls the type I error well, but at the expense of a slight decrease of power, which is expected because extra sources of error increase the random variation of the difference in allele frequencies between pools, so that a true difference is more difficult to detect. However, the power of the pooled sample can be maintained at a level equivalent to that obtained by individual genotypes by minimising the experimental error and slightly increasing the sample size.

Genotyping accuracy has not been systematically examined but recent studies (16; L. Peltonen, personal communication) have suggested that no genotyping method is 100% accurate, and that as many as 5% of individual genotypes could be miscalled. Such genotyping errors would decrease the power to detect quantitative trait loci (17) or could have serious effect on linkage disequilibrium measures (18). Most of the scoring errors are caused by ambiguities in the allele peaks, sample-to-sample contamination or mislabelling of DNAs. The use of pools should reduce all of these sources of error. Hence, the procedures described above can be used to perform very accurate association studies, saving valuable time and money compared with genotyping individual samples. Pooling studies should therefore be used to perform a fast, cheap and reliable preliminary screen of a candidate region.

ACKNOWLEDGEMENTS

We would particularly like to thank Christiane Honisch from Sequenom for leading the MassARRAY™ experiments, and Dirk van den Boom, Edvin N. Munk and Suzanne Müller, from Sequenom, for their support and comments. We would also like to thank Nadine Norton, Mick O'Donovan and Mike Owen (University of Wales College of Medicine) and Gerome Breen (University of Aberdeen) for sharing their useful expertise. We would like to thank Kenneth Humphreys from

the Gastro-Intestinal Unit, University of Edinburgh, and the MRC Human Genetics Unit, for use of their equipment. This work was supported by grants from the UK Medical Research Council, the UK Biotechnology and Biological Sciences Research Council and Organon NV.

REFERENCES

1. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Cardon,L.R. and Bell,J.I. (2001) Association study designs for complex diseases. *Nature Rev. Genet.*, **2**, 91–99.
4. Syvanen,A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.*, 2, 930–942.
5. Hoogendoorn,B., Norton,N., Kirov,G., Williams,N., Hamshere,M.L., Spurlock,G., Austin,J., Stephens,M.K., Buckland,P.R., Owen,M.J. and O'Donovan,M.C. (2000) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum. Genet.*, **107**, 488–493.
6. Breen,G., Harold,D., Ralston,S., Shaw,D. and St Clair,D. (2000) Determining SNP allele frequencies in DNA pools. *Biotechniques*, **28**, 464–466, 468, 470.
7. Germer,S., Holland,M.J. and Higuchi,R. (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.*, **10**, 258–266.
8. Sasaki,T., Tahira,T., Suzuki,A., Higasa,K., Kukita,Y., Baba,S. and Hayashi,K. (2001) Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am. J. Hum. Genet.*, **68**, 214–218.
9. Zhou,G., Kamahori,M., Okano,K., Chuan,G., Harada,K. and Kambara,H. (2001) Quantitative detection of single nucleotide polymorphisms for a pooled sample by a bioluminometric assay coupled with modified primer extension reactions (BAMPER). *Nucleic Acids Res.*, **29**, e93.
10. Buetow,K.H., Edmonson,M., MacDonald,R., Clifford,R., Yip,P., Kelley,J., Little,D.P., Strausberg,R., Koester,H., Cantor,C.R. and Braun,A. (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl Acad. Sci. USA*, **98**, 581–584.
11. Durbin,R. and Thierry Mieg,J. (1991) A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.
12. Hoogendoorn,B., Owen,M.J., Oefner,P.J., Williams,N., Austin,J. and O'Donovan,M.C. (1999) Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. *Hum. Genet.*, **104**, 89–93.
13. Sokal,R.R. and Rohlf,F.J. (1995) *Biometry*. WH Freeman and Company, New York.
14. Norton,N., Williams,N.G., Williams,H.J., Spurlock,J., Kirov,G., Morris,D.W., Hoogendoorn,B., Owen,M.J. and O'Donovan,M.C. (2002) Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.*, **110**, 471–478.
15. Evans,K.L., Le Hellard,S., Morris,S.W., Lawson,D., Whitton,C., Semple,C.A.M., Fantes,J.A., Torrance,H.S., Malloy,M.P., Maule,J.C., Humphray,S.J., Ross,M.T., Bentley,D.R., Muir,W.J., Blackwood,D.H.R. and Porteous,D.J. (2001) A 6.9Mb high resolution BAC/PAC contig of human 4p15.3-p16.1, a candidate region for bipolar affective disorder. *Genomics*, **71**, 315–323.
16. Bray,M.S., Boerwinkle,E. and Doris,P.A. (2001) High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum. Mutat.*, **17**, 296–304.
17. Abecasis,G.R., Cherny,S.S. and Cardon,L.R. (2001) The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.*, **9**, 130–134.
18. Akey,J.M., Zhang,K., Xiong,M., Doris,P. and Jin,L. (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am. J. Hum. Genet.*, **68**, 1447–1456.